Draft Report

Probabalistic Dilution Model 3 (PDM3)

Volume 1 of 2

Contract No. 68-02-4254
Task No. 117

Prepared by:

Versar Inc.
6850 Versar Center
Springfield, Virginia 22151

Prepared for:

U.S. Environmental Protection Agency
Office of Toxic Substances
401 M Street, S.W.
Washington, D. C. 20460

May 31, 1988

Disclaimer

This document is an interim report. It has not undergone a full technical and editorial review. The report has not been released formally by the Office of Toxic Substances, Office of Pesticides and Toxic Substances, U. S. Environmental Protection Agency. It is being circulated for initial comments on its technical merit and approach and policy implications.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Background

The USEPA Office of Toxic Substances, Exposure Evaluation Division (EED), is frequently required to perform exposure assessments on chemicals that may be discharged to streams by industrial facilities. Part of the assessment requires estimating the concentration of the chemical in the receiving stream. In the early stages of the new chemical review process, these estimates are based on the quantity of chemical released, the efficiency of wastewater treatment, and receiving stream flow rates.

The most important process affecting the concentration of a dissolved chemical once it is in surface waters is dilution and removal by advection. For screening level exposure assessments, EED will estimate the concentration of the chemical by simply accounting for its dilution i.e., (amount released) ÷ (stream flow). Recently, EED has begun using a modified computerized version of the USEPA's Office of Water Regulation and Standards Probabilistic Dilution Model (PDM) to help account for some of the variability that occurs with stream flows. The model is used to estimate flow variability of a reach using estimated flow values and coefficients of variation for hydrological basins. Additional calculations at the end of the model report the frequency a concentration of concern in a reach will be exceeded.

The model is based on simple stream dilution calculation:

$$C = \frac{L}{Q}$$

where

C = surface water concentration
L = chemical loading
Q = receiving stream flow rate.

The complexity of the model arises in its attempt to account for the natural variability of stream flows and effluent flows. In order to

account for this variability, a probability distribution for flow rates is incorporated into the above equation. The calculation of probability assumes that receiving stream flow, effluent flow, and effluent concentration are log-normally distributed and independent. The statistics involved include both the arithmetic and logarithmic forms of the mean, standard deviation, and coefficient of variation for the flow and concentration of both the stream and effluent.

A mathematical derivation of the calculation of the probability distribution is presented in DiToro (1984). Some applications of a special form of the model are discussed in the Technical Guide Manual for Performing Waste Load Allocations Book VII, Permit Averaging Period (USEPA 1984) prepared for the Office of Water Regulations and Standards. The special form of the model and a discussion of its mathematical derivation are presented in Section 5.

1.2 Purpose

In January 1987, the old version of PDM used by EED was improved by combining two separate but related programs into one package and, at the same time, making it more accurate, broader in scope, and user-friendly. The original EED programs were written in Basic. The January 1987 version was written in Pascal using the Basic version as the guide.

The new program was divided into two separate options, Option 1 for analysis of specific reaches and Option 2 for analysis of reaches related to specific industrial categories. They both use estimated flow data and employ the same probability of exceedance calculations except that the first option performs the calculation once and the second option performs it multiple times. Scope and accuracy of the program were increased by accounting for all hydrological regions on a more refined level (subbasin rather than basin). The model was made more user-friendly in-terms of the program packaging of the two different options, the user prompts and responses, and most importantly, by reducing the time necessary to run the model for the second option to a few seconds.

The January 1987 version used estimated flow values and probability calculations for all reaches regardless if actual flow data was available. During the summer of 1987, the program was further improved by adding a new feature to Option 1. This new feature allows the use of actual daily flow data for those reaches with USGS gaging stations. For these reaches, the distribution of daily flows can be determined directly without having to estimate it using probability calculations. In addition to this improvement, another improvement was made to the program: the estimated flow values used for reaches without gaging stations were added to the program so that the user does not have to retrieve them manually. This latest version of the program was referred to as PDM3. Option 1 - analysis of a single reach, thus has three possibilities, analysis for reaches with measured flow data, analysis for reaches with only estimated flow values, and analysis for reaches with user specified flows. The program will immediately indicate to the user whether a reach has actual flow data or not and select the appropriate analysis method.

PDM3 was then modified once again in November 1987. This modification was confined to Option 2 - analysis by SIC groupings. The previous version of Option 2 was an analysis for an average case probability. The modification changed the analysis from an average case to a worst case probability. This was accomplished using 10th percentile mean and low receiving stream flows for the SIC industrial groupings. The method for calculating the probability and the user interface for Option 2 were significantly different than the previous version of PDM3. Option 1 in this version remained the same.

A decision was made by EED in February 1988 to change the November 1987 version of PDM3. Two modifications were requested. First, to change the method by which the worst case probability was being calculated in Option 2. The second was to change some of the user interface features of the program. These interface changes affect both Options of the program.

This report serves as the documentation for PDM3. It replaces the reports created under Task 16 and Task 37 of this contract; much of those

reports are contained here unchanged. The report is in two volumes and has been modified to reflect the changes since the August 1987 and November 1987 versions. In addition to this report, a revised user's manual providing easy to follow example sessions of PDM3 has been prepared under a separate cover. And, under another separate cover, a programmer's documentation outlining both the mainframe and PC programs has been prepared for future reference; copies of the source codes are contained in the report.

The remainder of Section 1 will further describe PDM3, its specific uses (options), and the methods used to improve it from the old (1986) EED version. Section 2 describes the data and data sources necessary to run PDM3. Sections 3, 4, and 5 document the development of Option 1. Section 3 describes the analysis for reaches with measured daily flows. Sections 4 and 5 describe the development of the probability calculations used for reaches with only estimated flow data or user specified flow data. Section 6 is documentation for the development of Option 2 of the program.

1.3 Explanation of PDM3 Options

The PC package described in this report is designed to model the needs of EED. The package is divided into two approaches (Options) either of which the user may select depending on the extent of data available. The first approach is when the user knows the location of the discharger or point source and can identify the receiving stream or would like to analyze a hypothetical reach. The second case is when a specific site of a source is not known or when a generic case is being examined and an analysis of an industrial category is necessary.

1.3.1 Option 1

The first approach addresses site-specific cases. For the typical case, the user will usually know the following prior to using PDM3:

- Name, NPDES number, and location of industrial plant;
- Receiving stream 11-digit reach number;

- Effluent flow rate from the plant;
- Number of release days per year.

The user will input the reach number of the river segment to be analyzed. The program will then determine whether the reach has actual, estimated, or no flow data and then guide the user accordingly. The latter case are for reaches that are lakes, estuaries, bays, etc.

The user will then be prompted for the number of release days, loading, and a concentration of concern (COC) level. The model will then determine the frequency (number of times per year) the receiving stream's concentration will exceed the concern level. The program is designed such that the user can easily vary the release days, loadings, and COC level inputs to produce a series of results. The last subsection of Section 5 describes the calculations for the final product of Option 1.

One of the improvements to the model involves the statistics of the receiving stream flows. Heretofore, stream flow variance was calculated on a USGS Hydrological Basin basis. There are 18 USGS basins in the contiguous United States. As part of this work, stream flow variance was analyzed on a subbasin level; each basin is divided into subbasins. The number of subbasins is dependent on the size and hydrogeography of the basin; there are 204 subbasins. Since the area of a subbasin is much smaller than its parent basin, an analysis of flow variance of this level will yield a more accurate representation of an individual stream.

Coefficients of variation for stream flows were calculated for each subbasin using standard statistical analyses, mean and standard deviations, and regression analyses. The coefficients are used in the probability calculations to account for the flow variability that might occur in a particular stream. Section 4 describes in detail the analyses used to develop the coefficients of variation for each subbasin.

If a hypothetical reach is to be analyzed, such that the user can input his/her own reach flows, then reach "0" may be entered.

1.3.2 Option 2

The second approach involves cases where the location of the chemical loading is unknown. In this case, the user will have to know the use of the chemical substance in order to select an SIC representative of the industry since stream flow data are grouped according to SIC. The user will then have to enter the number of release days per facility, the chemical loading, and the concentration of concern to be modeled. Data files have been stored in the program according to industrial classifications. These stored data are accessed by the PC program to calculate the frequency of exceedance.

This new version has three distinct calculation improvements from the old (1986) EED version:

1. The receiving streams for both direct and indirect dischargers are accounted for;

2. All the receiving streams in a SIC category are used to determine the probability and not just every representative fifth percentile flow (i.e., 5th, 10th, 15th, ...);

3. The appropriate subbasin coefficients of variation are applied to each stream flow for calculating probability. No coefficient was previously applied; and

4. The analysis is a worst case scenario providing answers appropriate to EED's exposure assessment needs.

Finally, the running time of this option has been reduced to several seconds rather than minutes or even hours. The improvements regarding the probability calculations are accomplished by an extensive effort in analyzing the data in the Hydrologically Linked Data Files (HLDF) maintained by OWRS. The reduction in running time is accomplished by running the probability calculations for each SIC group thousands of times on the EPA IBM 3090 mainframe computer to create a matrix file of probabilities. These matrix files are stored on a PC diskette and are accessed based on the user's input. Thus, the PC program actually

interpolates the matrix files to determine the probability of exceedance. The interpolation is performed in a matter of seconds.

The methods used to accomplish the above improvements are outlined and discussed in detail in Section 6.

2. DATA DESCRIPTION

For a clear understanding of the methods used to prepare the PDM3 PC package, it is necessary to define some of the data elements used and their sources. As mentioned previously, the primary sources of data are the Hydrologically Linked Data Files (HLDF) and the STORET Flow File maintained by OWRS. This section describes the pertinent data files as well as data terms used throughout PDM3.

2.1 HLDF

The HLDF system is composed of the following five files:

Reach File;
GAGE File;
Industrial Facility Discharge (IFD) File;
Water Supply Data Base (WSDB); and
FISHKILL.

The files are hydrologically linked by an 11-digit number, referred to as a reach number, assigned to a particular surface water body, usually a river or river segment. Each of the above files contains particular information pertaining to a reach (river segment) and thus, the reach number is a common data element linking the files.

The USGS has divided the United States into major regions (river basins) which are further divided into subregions (subbasins), then accounting units, and finally, cataloging units. The EPA further divides cataloging units by assigning a 3-digit number to the major rivers, streams, and lakes that are in each cataloging unit. Thus, a reach number is composed as shown in the following:

Reach number

	03 05 01 02 009
Basin	--
Subbasin	-----
Accounting Unit	-------
Cataloging Unit	-----------

--- EPA segment number

Essentially, there are two types of reaches: river segments and shorelines of lakes, reservoirs, bays, and coastal waters. Reaches usually extend from one segment junction to another.

The first three of the above listed files are important to the PDM3 model developed here. The following is a brief discussion of these three files.

2.1.1 Reach File

This file contains the list of all reaches in the contiguous United States and their identifying numbers. There are approximately 72,000 segments in the 2,111 USGS hydrological cataloging units. Information in this file for each reach includes its latitude-longitude coordinates, segment name, downstream and upstream segments, length, and type of surface water body.

2.1.2 GAGE File

The GAGE data file contains information on approximately 37,000 stream gaging locations throughout the U.S. Information stored includes location of gaging stations, corresponding reach, types of data collected, frequency of data collection, media in which data are stored, identification of the collecting agency, and, where available, mean annual flow, 7-day-10-year (7-Q-10) low flow, and flow velocity. Reaches without a gaging station have estimated mean and low flows and velocities. Estimated values are indicated by the identifiers WEG and GKY; actual flow data are identified by the identifier USGS.

2.1.3 Industrial Facilities Discharge (IFD) File

IFD is a comprehensive data base of point source dischargers. Currently, there are more than 40,000 direct discharge facilities in the IFD file, of which nearly half are Publicly-Owned Treatment Works (POTWs). Included as contributors to POTWs are approximately 11,500 indirect discharge facilities.

There are approximately 130 data elements within the IFD data file for which data have been collected from various sources; the following sources are the most important:

1. Permit Compliance System (PCS);
2. EPA's 1978 NEEDS Survey File; and
3. NPDES permit files.

IFD is constantly being updated because of the ever changing data concerning industrial dischargers.

The IFD data file is organized as a hierarchical information system of three levels: facility level, discharge pipe level, and contributing indirect facility level. This organizational structure allows facilities to be viewed in their entirety or as separate discharge pipes within a facility. The facility level contains identification codes and summarized discharge information (e.g., name, address, NPDES number, city, county, total facility flow, SIC codes, and the receiving water reach number and name). The discharge pipe level includes the components of each individual discharge, such as location, flow, and SIC code activity. The indirect facility level includes data on industrial flow from industries that discharge to another facility, such as a POTW, rather than directly to surface water.

2.2 STORET Flow File

The Flow File is a collection of daily flow data gathered at USGS gaging stations throughout the country. USGS supplies new flow data into the file on a biannual basis; the file currently contains data up to August 1986. The file is composed of stream flow data indicating the quantity of water flowing past the gaging sites.

2.3 USGS Hydrological Basins

As described above, the USGS has divided the contiguous U.S. into 18 hydrological river basins or watersheds. These major basins are further divided into 204 subbasins. Table 2-1 lists the names of the basins and the number of subbasins in each basin.

Table 2-1. Name of Basins and Number of Subbasins

Basin No.	Basin Name	Number of Subbasins
01	Northeast	11
02	Mid-Atlantic	8
03	South Atlantic	18
04	Great Lakes Region	15
05	Ohio	14
06	Tennessee	4
07	Upper Mississippi	14
08	Lower Mississippi	9
09	Souris-Red-Rainy	3
10	Missouri	30
11	Arkansas-White-Red	14
12	Texas-Gulf	11
13	Rio Grande	9
14	Upper Colorado	8
15	Lower Colorado	8
16	Great Basin	6
17	Pacific Northwest	12
18	California	10

2.4 Flow Data

There are three different flow data items used in PDM. Two of the three deal with the receiving stream, and the third is the effluent flow rate from an industrial plant. The stream flow data are taken from the GAGE file and are measured flows recorded at 4,235 USGS gaging stations within the 204 subbasins. The effluent flow data are obtained from IFD.

Receiving Stream Mean Flow - The average daily flow calculated over the period of record to 1983 for each gaging station.

Receiving Stream Low Flow - The 7-Q-10 flow - the lowest recorded average flow over a 7-day period for a 10-year period at each gaging station.

Plant Flow - The flow reported on the industrial facility National Pollutant Discharge Elimination System (NPDES) permit.

The receiving stream flow data used in this study is from a previous study. Versar (1984) provides a complete description of the flow data retrieved from the GAGE file and the statistical analysis performed on it. There is an additional parameter that arises from the above flow data and used in the probability calculations, the Mean Flow Dilution Factor. This parameter is derived by dividing the receiving stream mean flow rate by the plant effluent flow rate. The resulting value is the number of times the effluent will be diluted once it is in the receiving stream.

2.5 Industrial SIC Groups

As previously mentioned, IFD contains information on industrial plants that discharge to surface waters, i.e., facilities that have NPDES permits. In addition, IFD also contains a small fraction of the industrial plants that discharge to POTWs. All of these facilities are listed in IFD with their respective SIC codes. The exposure assessments performed by EED do not involve many SIC categories. Therefore, only facilities with certain SIC codes were retrieved from IFD and their information used in PDM3. There are 39 groups that EED usually deals

with. Table 2-2 lists these SIC groups. Some of the groups have multiple SIC codes since these industries are closely related and are likely to be using the same chemicals and chemical processes. Data was retrieved and analyzed for each group and used for Option 2 of PDM3.

2.6 Data Problems

The HLDF system is a large and ever changing data base and as such is difficult to maintain. In particular, the IFD file will consistently be reporting old data or be missing data because of the ever changing status of NPDES permitted dischargers. Some of the more common data problems within the file are:

- Facilities with bad SIC codes;
- Facilities without identified pipe types (C, P, or B);
- Facilities with multiple pipes with some missing data;
- Facilities without effluent flow data;
- Facilities with incomplete, missing, or wrong reach numbers; and
- POTWs with only a partial list or no listing of their industrial dischargers.

Even with these problems, IFD is the most comprehensive, accurate and available data source on industrial dischargers. For the purpose of PDM3, only those facilities with a complete set of data necessary for the PDM calculation were used. A further description of these facilities is in Section 6.1.

Table 2-2. Industrial Groupings in PDM3

SIC Codes	Industry
2891	Adhesives and Sealants Manufacture 50
3674, 3679	Electronic Components Manufacture
3411	Metal Can Manufacture
3471	Electroplating
332, 336	Foundries
2865, 2869	Organic Chemicals Manufacture 246
2893	Ink Formulation
281	Inorganic Chemicals Manufacture
3111	Leather Tanning & Finishing
2911, 2992	Lubricant Manufacturers
(See * Below)	Metal Finishing
2851	Paint Formulation
101-109	Ore Mining & Dressing
2621, 2631, 2661	Paper and Paperboard Mills
2621	Paper Mills, except Building Paper Mills
2631	Paperboard Mills
2661	Building Paper and Board Mills
2819, 2869, 2879	Pesticides Manufacture 451
2911	Petroleum Refining
7221, 7333, 7395, 7819	Photographic Processing
3079	Plastic Products Manufacture
2821, 2823, 2824	Plastic Resins and Synthetic Fibers Manuf.
(See ** Below)	POTWs (Industrial)
4952	POTWs (All) 1026
271-277	Printing
2611	Pulp Mills
3011, 3021, 3031, 3041	Rubber Products Manufacture
2841, 2842, 2843, 2844	Soaps, Detergents, etc. Manufacture
7211, 7213-7219, 7542	Auto and other Laundries
3711, 3713	Motor Vehicle Manufacture
3631, 3632, 3633, 3639, 3431, 3469	Large Household Appliance and Parts Manufacture
3315-3317, 3351-3357, 3463, 3497	Primary Metal Forming
2281, 2282, 2283, 2284	Yarn and Thread Mills
2271, 2272, 2279	Carpet Dyeing and Finishing
2231	Wool Dyeing and Finishing
225, 2292	Knit Fabric Dyeing and Finishing
2261, 2262, 2269	Woven Fabric Dyeing and Finishing
2231, 225, 226, 2292	All Textile Dyeing, except Carpets
4911	Steam Electric Plants

* Metal Finishing = 3411-3462, 3465-3471, 3482-3599, 3613-3623, 3629, 3634-3636, 3643-3651, 3661-3671, 3673, 3676-3678, 3693-3694, 3699, 3711-3841, 3851, 3873-3999

** POTWs (Industrialized) = IDSI = 1011-1999, 2211-5199, 5511-5599, 7211-8099

3. OPTION 1 - REACHES WITH USGS GAGING STATIONS

As mentioned in Section 1, Option 1 of PDM3 is used to analyze a specific reach. Approximately 1,500 of the approximately 37,000 EPA identified river segments (reaches) have USGS flow gaging stations located on them and thus, measured daily flow values are available for them. The remaining reaches have no measured flow values that are readily accessible, only mean and low flow estimates are available for them. Obviously the measured flow data is the more preferable data to use when trying to determine the frequency of a concentration of concern (COC) being exceeded. That is why an additional feature was added to Option 1, so that a more accurate assessment could be made for those reaches with actual data.

This section describes the development of the program which references stream flow data measured from USGS gaging stations over a twenty year period. This data was formulated by first accessing the Stream Gaging Inventory File (GAGE), which is one of the five data files comprising tne Hydrologically Linked Data Files maintained by OWRS, and then the Storet Flow File. The first two subsections in this section briefly describe the use of these two files and the data elements extracted from them. The third subsection describes the creation of the PC files used in PDM3. The fourth subsection describes the development of the PC program and the final calculations. Figure 3-1 is a flow chart depicting the development of the program.

3.1 Stream Gaging Inventory File

As mentioned in Section 2, the Gage File contains mean and low flow data on river segments (reaches) that have been assigned a USEPA reach number. Some of the data are statistical averages from measured data from USGS gaging stations. However, most reaches only have estimated mean and low flow values since no gaging station is located on them. The big advantage of the Gage File is that it identifies gaging stations to their respective reaches. The Gage File was used to create a listing of

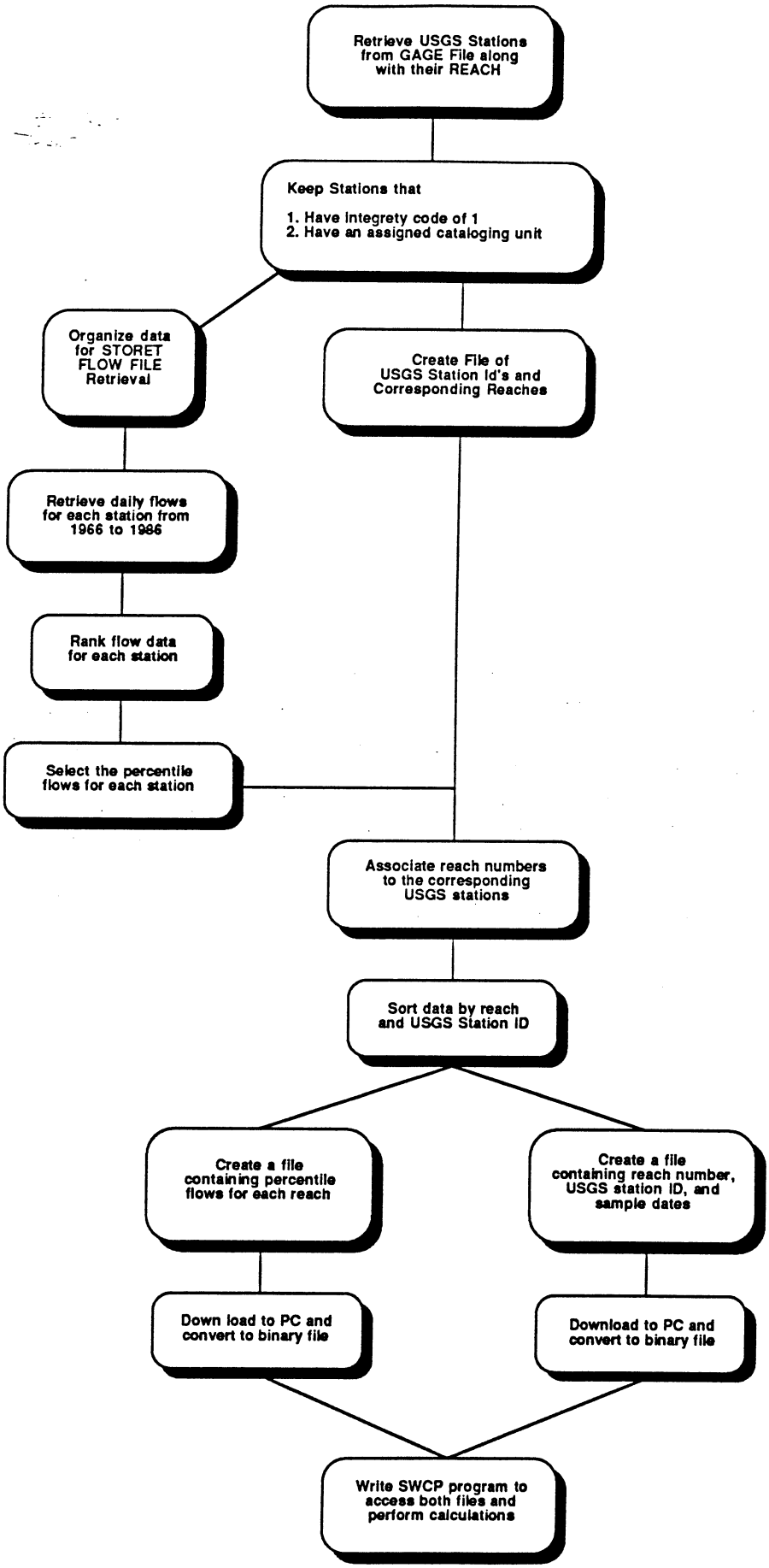

Figure 3-1. Flow diagram for development of program for reach with gaging stations.

all USGS gaging stations that are located directly on a reach. To do this, all stations were retrieved with the following data elements:

- State/County FIPS Code
- Agency Code and Gage ID
- Integrity Code
- Cataloging Unit
- Segment Number

The information was placed in a temporary mainframe file. This file was then edited to remove those reaches that did not have a gaging station located on them. An integrity code with a value of one indicates that the station is located on the reach. Therefore any station with an integrity code other than one was removed from the temporary file. The data was then formatted to make it compatible with the Storet Flow File retrieval program Flow and saved in another file. The in-house software of OWRS was used to retrieve, edit, and format the data.

3.2 STORET Flow File

Daily flow data was retrieved from the file for gaging stations by specifying the state FIPS code preceded with an 'S' and followed by the station ID. For the purpose of PDM3, the listing of gaging stations retrieved from the Gage file was divided into several datasets and passed to the Flow file retrieval program FLOW as include files. FLOW was directed to report all daily flow values for each station for the years 1966 through 1986. The output from the Flow file was then directed to a SAS program step which reads the data into a format suitable for manipulation. The total number of daily flow values is also calculated and the period of record is also identified.

Flow data was not found for all of the stations passed to the FLOW retrieval program. There are two reasons for this. First, the station was present within the Storet Flow File but there was no data for the specified time period, or second, the station was not found in the Flow file (data for these stations must have originated from data bases other than STORET).

3.3 Data Files

The data used in this program originated from the GAGE and FLOW file retrievals mentioned above. A SAS program was written to rank the daily flow data for each station and then choose 100 percentile flows. Another SAS dataset was created from the data retrieved from the GAGE file which contained the station ID and its associated reach. The data was ranked using the SAS procedure PROC SORT which utilizes the SyncSort OS software product. The percentile flows were chosen from the ranked list of daily stream flows for each gaging station by picking every Ith observation such that I = Int(n*k/100) where k varies from 1 to 100 with each percentile flow picked and n is the number of daily flow values retrieved for the gaging station and the Int() function returns the largest integer value equal to or less than the value of its argument.

Because percents are picked in this way, a gaging station with less than 100 daily flow values will have no percentile flows calculated. For example, if n = 60 and k = 1 (starting condition) then I will have the value of 0. Since SAS datasets begin with observation number one, no percentile flow is chosen and k is not incremented. Thus, a gaging station in the Flow File with less than 100 data values will not be found as a reach with a gaging station.

The dataset containing the percentile flows was then merged with the dataset originating from the GAGE file to associate gaging stations with their corresponding reaches. Finally the data was sorted in ascending order according to reaches and station ID.

In preparation for data downloading the final SAS dataset was divided into two parts and written as sequential OS datasets. The resulting datasets are referred to as the "flow file" which contains only flow data and the "reference file" which contains the reach number, station ID, and flow data period of record information.

3.4 Development of PC Program

3.4.1 Data File Downloading

The data files were transferred to PC diskettes from the mainframe using a TSO file transfer protocol. To ensure complete transfer, the PC files were compared to the sequential files stored on the mainframe.

3.4.2 Conversion of Data From ASCII to Binary Format

Two programs were written in Pascal to read the downloaded data and to put it in special record formats used by PDM3. The data was then written to new files in a binary format suitable for random access. The new files were checked to ensure that they contained the same data in the same order as the original downloaded ASCII files.

3.4.3 Final Calculations

A PC program to access the data files and produce the final results was written in Pascal. A copy of the program is in Appendix J. The major parts of the program are:

1. Determine if the reach has a gaging station on it.
2. Access the 100 percentile flows for the reach.
3. Report the station ID number, number of observations, period of record, 50th and 10th percentile flows, and number of stations on the reach.
4. Prompt the user for number of release days and loading rate, and then COC.
5. Calculate "Percent of Year Exceeded" and "Days/Year Exceeded" and report the results.

The above calculations are performed as follows:

1. Calculate the concentration for each percentile flow by dividing the loading rate by the percentile flow.
2. Compare the COC with the 100 calculated concentrations and select the highest percentile flow that yields a concentration greater than the COC.

3. Calculate Percent of year exceeded by the following:

$$\text{Percent of year Exceeded} = \frac{(\text{Release Days})(\text{Percentile from above})}{(365 \text{ days/yr})}$$

4. Calculate Days per year exceeded by the following:

$$\text{Days/yr Exceeded} = (\text{Release Days})(\text{Percentile from above})$$

When using this option, the user should take note of the number of observations reported for the station. Some stations may only have a few hundred flow values and thus may not be sufficient for this analysis.

Finally, some reaches have multiple gaging stations on them. The program will indicate this and the user must select which to perform the analysis. Unless the user knows the locations of the stations on the reach, it might be advisable to run the COC for each station.

4. OPTION 1 - REACHES WITHOUT GAGING STATIONS - CHARACTERIZATION OF UPSTREAM FLOW

As mentioned in Section 1, the original version of the program used by EED employed probabilistic calculations presented in DiToro (1984) and USEPA (1984) to account for variability in stream flows for those reaches with only estimated mean and low flows. The coefficient of variation for flow is a primary input in the probability calculations. In the original program, these coefficients were developed on a USGS hydrological basin scale (Versar 1984). In an effort to improve the results, coefficients of variation were determined for each USGS hydrological subbasin.

The probabilistic calculations, presented in Section 5, show the relationship between the input variables, upstream flow (QS), effluent flow (QE), effluent concentration (CE) and the output variable, downstream concentration. As discussed below in Section 4-1, each of the input variables is characterized by its mean and its coefficient of variation. A calculation of the exceedance probability requires a knowledge of these means and coefficients of variation. Estimated values of the means and coefficients of variations can be obtained from available data (if any) on these variables. A determination of the coefficient of variation of upstream flow is presented in this section.

4.1 Analysis of Upstream Flow Data by Basins

In one of the early versions of the PDM, upstream flows were stratified by the USGS hydrological basins (18 basins). Data retrieved in Versar (1984) were analyzed for each basin. The analysis was performed using the following unitless variables:

x = ratio of 7-Q-10 upstream flow to the mean flow

y = coefficient of variation of upstream flow (ratio of the standard deviation of upstream flow to the upstream mean flow)

A statistical regression analysis was performed on the upstream flow data for each basin separately (Versar 1984). The analysis showed that

in 15 basins (basins 1-7, 9-12, 14, 15, 17 and 18), the coefficient of variation of upstream flow, y, is exponentially decreasing as the ratio of the 7-Q-10 to the mean flow, x, is increasing. This exponential decay can be described by the regression model $y = e^{A-BX}$, where A and B are the regression coefficients (A, B > 0). The exponential regression of y on x is equivalent to a linear regression of the natural logarithm of y, ln(y), on x. The correlation coefficients between the transformed variable ln(y) and x for these 15 basins were low (but significantly different from zero). The analysis of the data for basins 9, 13 and 16 showed no significant correlation between the transformed variable, ln(y), and x. The coefficient of variation of upstream flow in each of these three basins was determined to be a constant of the form, $y = e^{A}$, where A is the natural logarithm of the average of y values in the basin data. The values of the coefficients, A and B or A only, were computed for each basin. In the execution of the PDM, a value x for the basin is computed as the ratio of the user input values of the 7-Q-10 upstream flow and upstream mean flow. This value, x, is then used along with the corresponding A and B values of the user's specified basin to compute the coefficient of variation of the upstream flow using the equation $y = e^{A-Bx}$.

4.2 Analysis of Upstream Flow Data by Subbasins

The analysis of the upstream flow data by basins (Versar 1984) showed that the values of the percentages of variations for the transformed variable, ln(y), explained by its linear regression on x, termed the R^2 value (the squared value of the correlation of coefficient), were relatively low. The statistical characteristics of the upstream flow rates differed among the subbasins in each basin, and, hence, the combined data of the subbasins within a basin would tend to have a higher variability than the variability of the subbasin data sets separately. Therefore, it was determined to study the characteristics of these upstream flow data by subbasins. The data were stratified by subbasins. The number of subbasins in each basin is shown in Table 2-1. The number

of data points, range, mean, and standard deviation for each subbasin are presented in Appendix A. An analysis of covariance was used to investigate the statistical significance of the differences between subbasins within a basin, a scattergram of the data by subbasins were used to study the patterns of relationships between y and x in each basin and a regression analysis was performed to determine these regression relationships.

4.2.1 Analysis of Covariance of Upstream Flow Data by Subbasins

To investigate the significance of the differences between subbasins within a basin, a comparison of the adjusted means of the transformed variables ln(y) (adjusted for their regression on the variable x) was performed for each basin. The comparison was performed through an analysis of covariance (ANCOVA). The results of the ANCOVA for each basin are listed in Appendix B. Each table in Appendix B shows the result of the test of equality of the adjusted subbasin means compared with the adjusted mean for the entire basin. The test statistic is provided by the number labeled "PR-F" (probability of exceeding the corresponding F value). This test statistic indicates the probability of no significant differences among the adjusted means of the subbasins. The PR-F value is smaller than 0.01 for Basins 2 through 13 and 15 through 18, smaller than 0.05 for Basin 1, and smaller than 0.10 for Basin 14. Thus, at a significance level of 0.05, the ANCOVA results showed significant differences between the adjusted means of the transformed data for the subbasins in each basin except Basin 14. Therefore, data for subbasins in each basin should be used to develop separate subbasin coefficient of variation values. Only if there are limited data or other technical constraints should all of the basin-wide data be combined for an analysis of the coefficient of variation for the entire basin.

4.2.2 Scattergrams of the Upstream Data by Subbasin

To examine the relationship between y and x, the y values were plotted versus the x values for each subbasin. Some of the subbasins

were found to contain gaging stations that resulted in "outlier" data. An outlier among the y values is one that is far greater or smaller than the rest and lies three standard deviations or further from the mean of the y values. The outlier is a peculiarity and indicates a data point that is not at all typical of the rest of the hydrologic data for that subbasin.

Examples of the various types of scattergrams are shown in Figures 4-1 through 4-5 at the end of this section. The patterns shown in the scattergrams are the following:

Pattern 1. Graphs showing an upward curvature without outliers (Figure 4-1).

Pattern 2. Graphs showing an upward curvature, but with outliers (Figure 4-2).

Pattern 3. Graphs showing a random scatter (a horizontal band) (Figure 4-3).

Pattern 4. Graphs showing that the majority of the points are located at the zero-value of the x axis (Figure 4-4).

Pattern 5. Graphs of subbasins with few gauging stations and hence only a few data points (one, two, or three) (Figure 4-5).

The majority of the subbasins had scattergrams similar to those of Pattern 1 or Pattern 2. The scattergrams of Pattern 1 indicate that an exponential model* of the form

$$y = e^{A-Bx}$$

would fit the regression of y on x.

---

* The use of the power model was investigated, but the exponential model outperformed the power model in terms of the R^2 and mean square error.

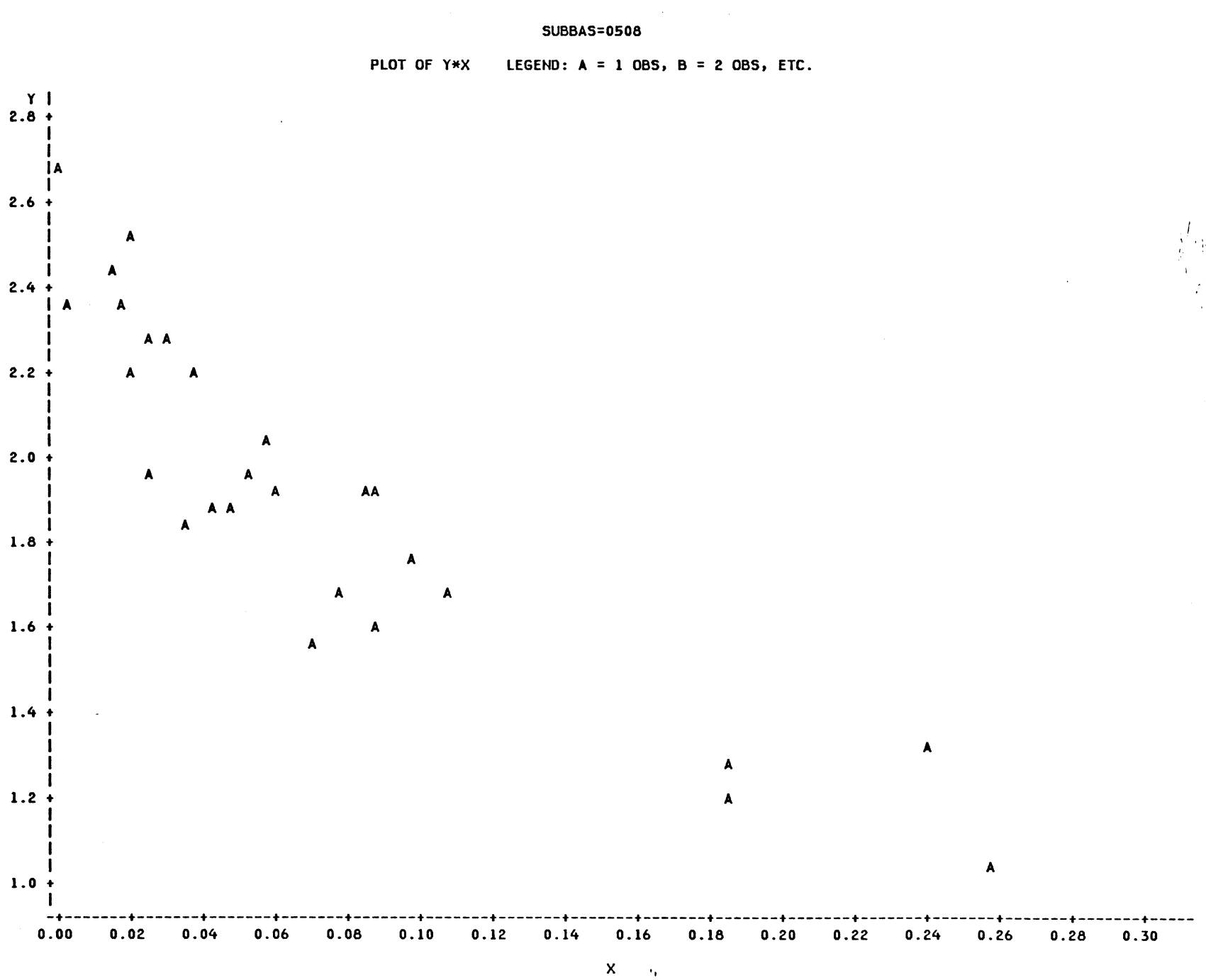

Figure 4-1. Data for Subbasin 0508 showing an upward curvature without outliers.

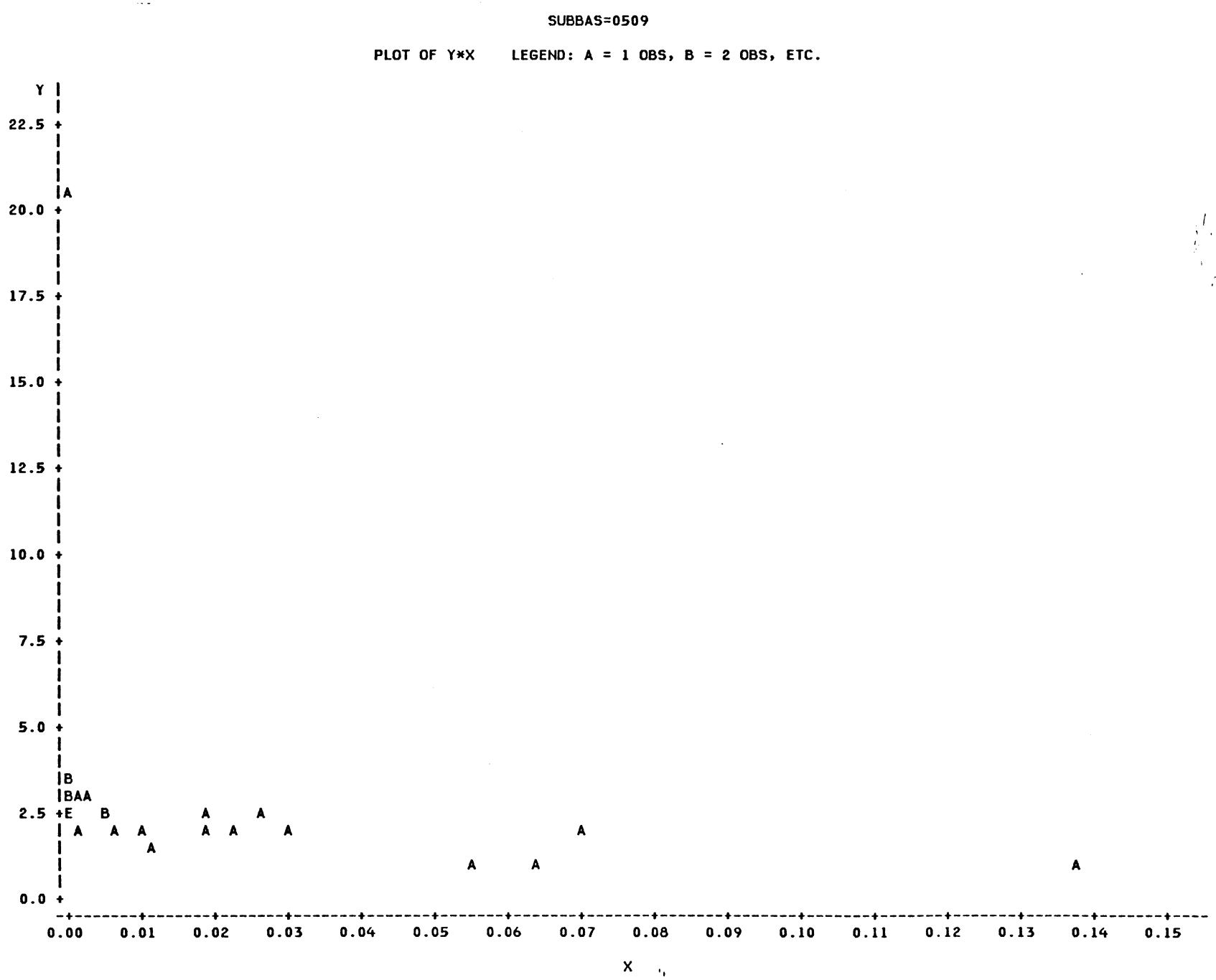

Figure 4-2. Data for Subbasin 0509 showing an upward curvature, but with outliers.

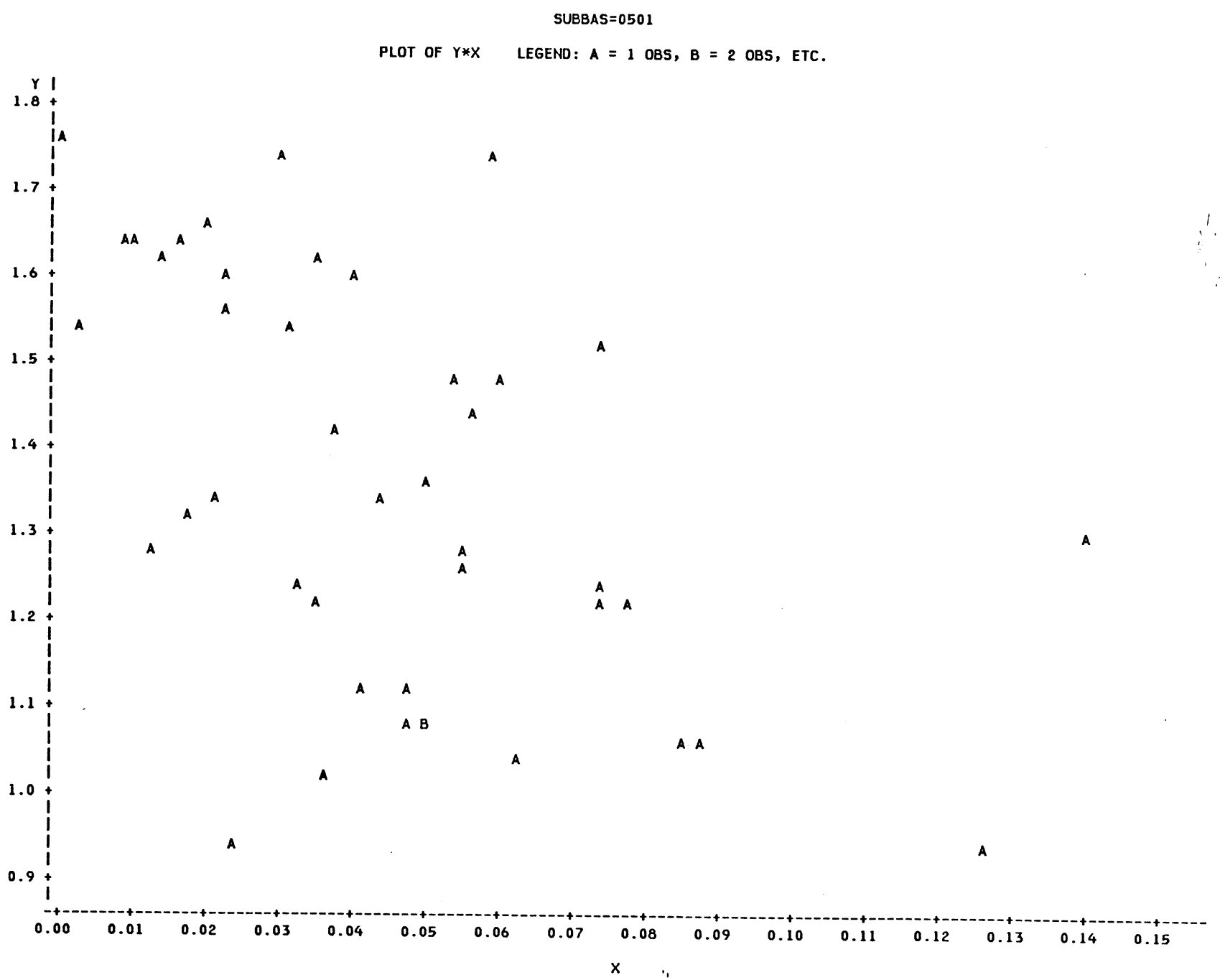

Figure 4-3. Data for Subbasin 0501 showing a random scatter of the data.

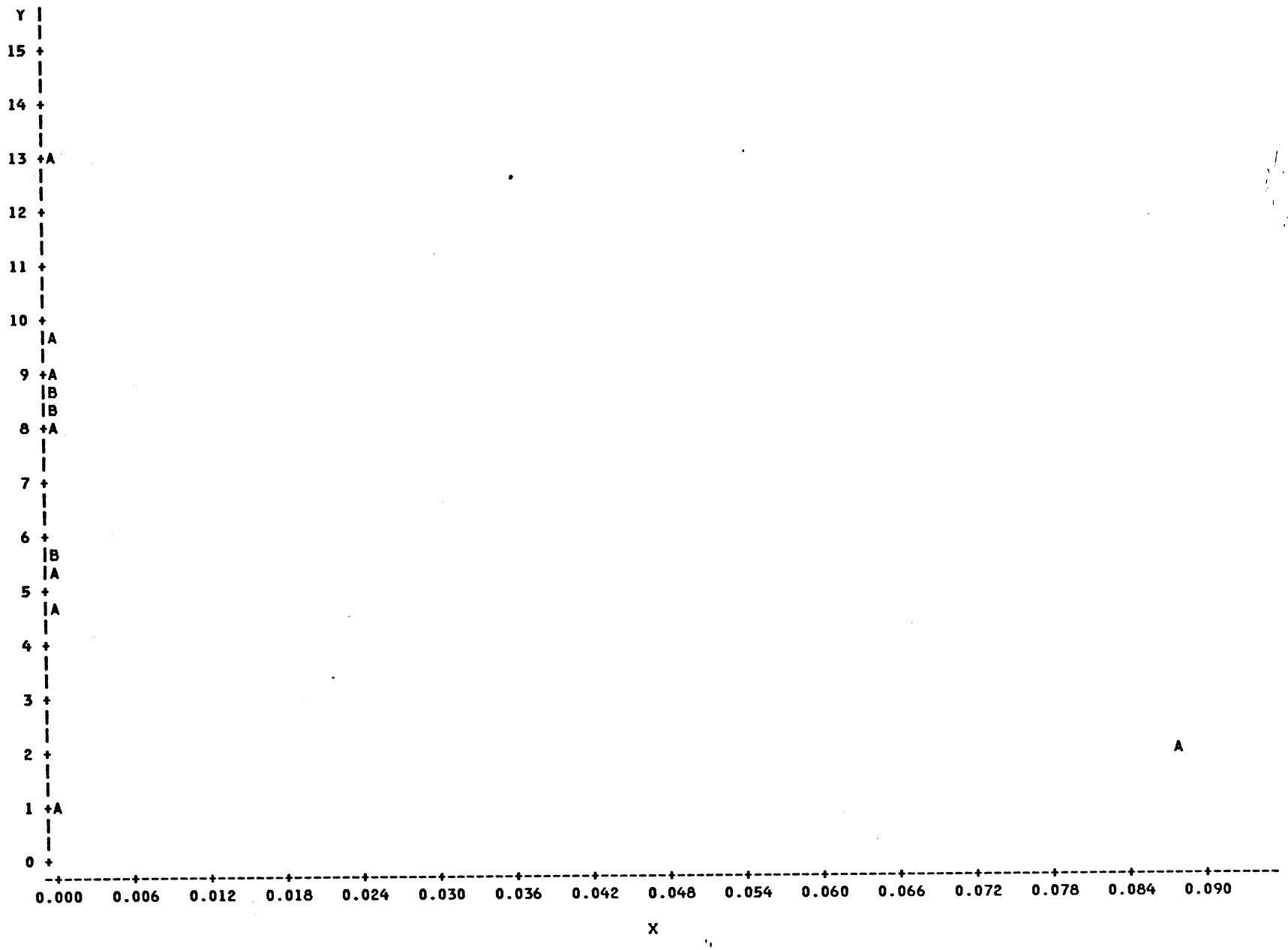

Figure 4-4. Data for Subbasin 1807 showing the majority of the data located at the zero value of the x axis.

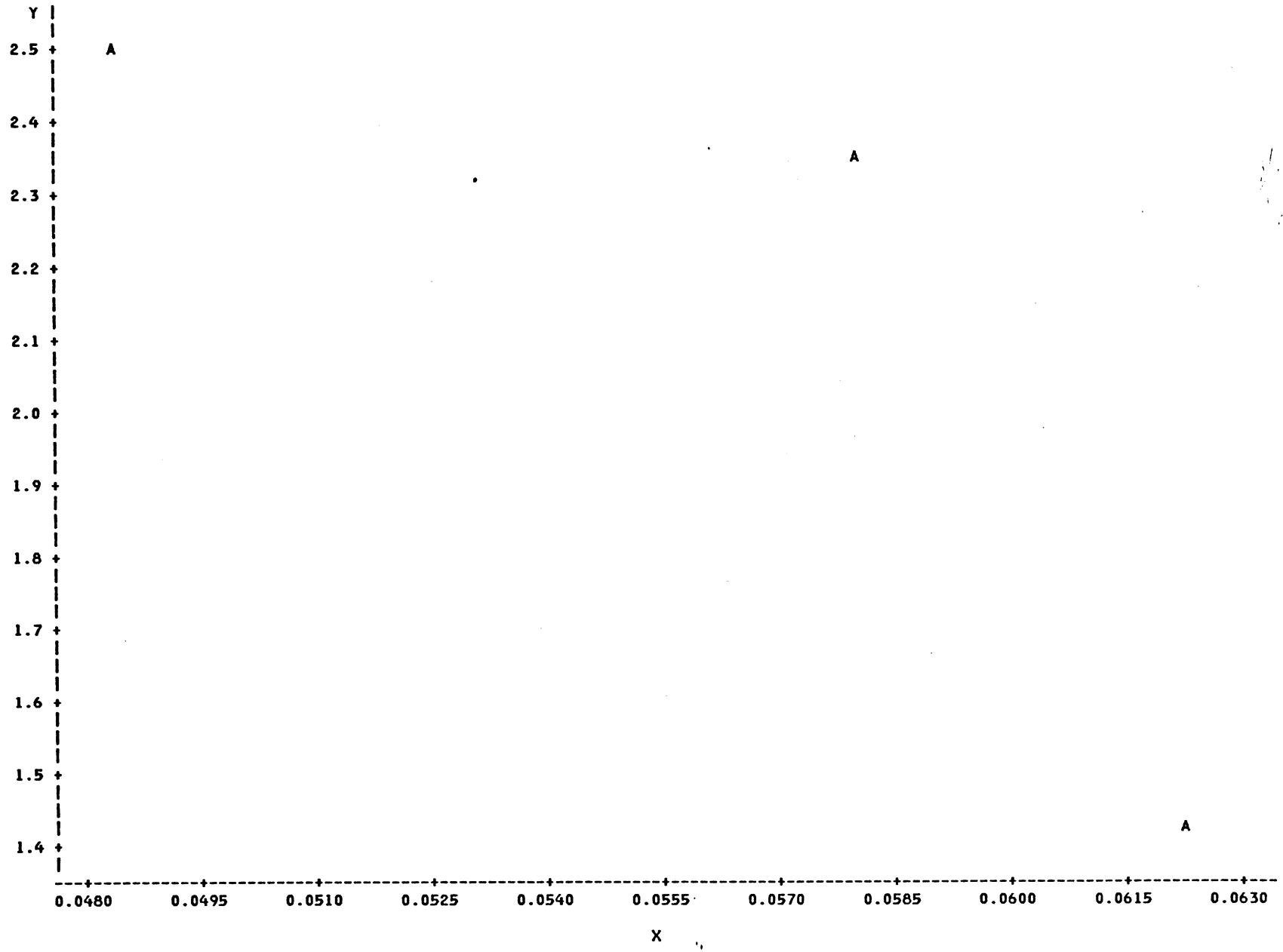

Figure 4-5. Data for Subbasin 0302 showing a subbasin with only a few data points.

The scattergrams of Pattern 2 indicate that the outlier(s) should be truncated and then the exponential model would fit the remainder of the data.

The scattergram of Pattern 3 indicates that the data do not show any relationship between x and y, and hence the predicated value of y is independent of the x values and the best predictor of the ln(y) value is the average value.

The scattergram of Pattern 4 indicates a similar conclusion to that of Pattern 3, that the best predictor of ln(y) is the average of the ln(y) values.

The scattergram of Pattern 5 indicates that the relationship between x and y cannot be inferred from this small number of data points.

The scattergram in Figures 4-1, 4-2, and 4-3 represent the data for three subbasins (0508, 0509, and 0501, respectively) that belong to Basin 05. The data for each of the other subbasins in Basin 05 exhibit similar patterns to that of Figure 4-1. However, the curvatures of the graphs differ among each of these other subbasins. These differences are a result of the variance in the hydrological characteristics of each of the subbasins. This variance among subbasins is accounted for by the regression coefficients A and B of the exponential model for each subbasin. The above example typifies the relationship among all subbasins, not just those in Basin 05.

The natural log transformation of the coefficient of variation, y, was performed, and a scattergram of the transformed data, ln(y), versus x was plotted for each subbasin. A scattergram of the fitted values from the linear model

$$\ln(y) = A - Bx$$

versus the x values was overlayed on the graph of ln(y) versus x for each subbasin. Examples of these scattergrams are shown in Figures 4-6 and 4-7. Figure 4-6 represents the data for subbasin 0508, Pattern 1, and

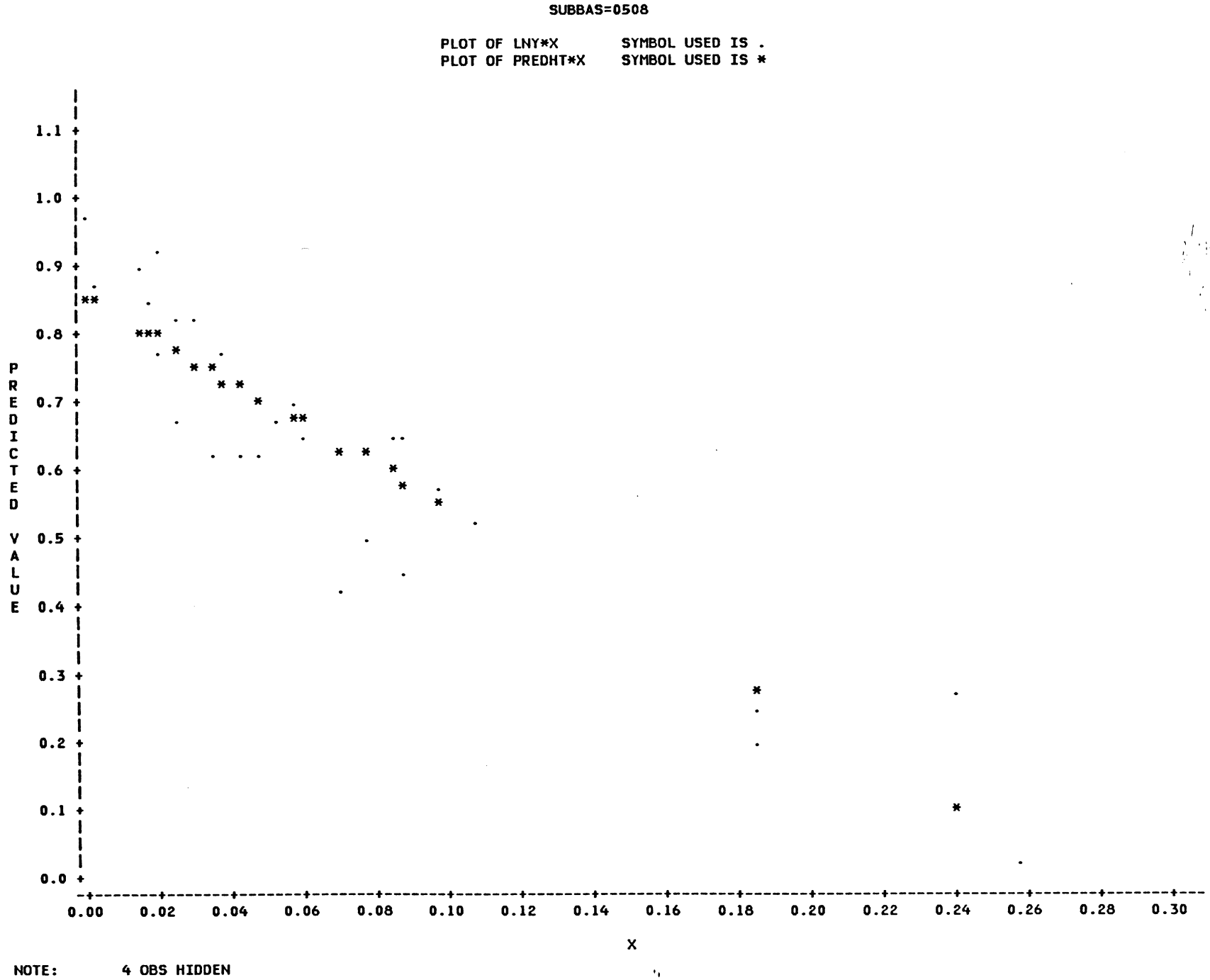

Figure 4-6. Actual data and predicted exponential equation for Subbasin 0508 (Pattern 1).

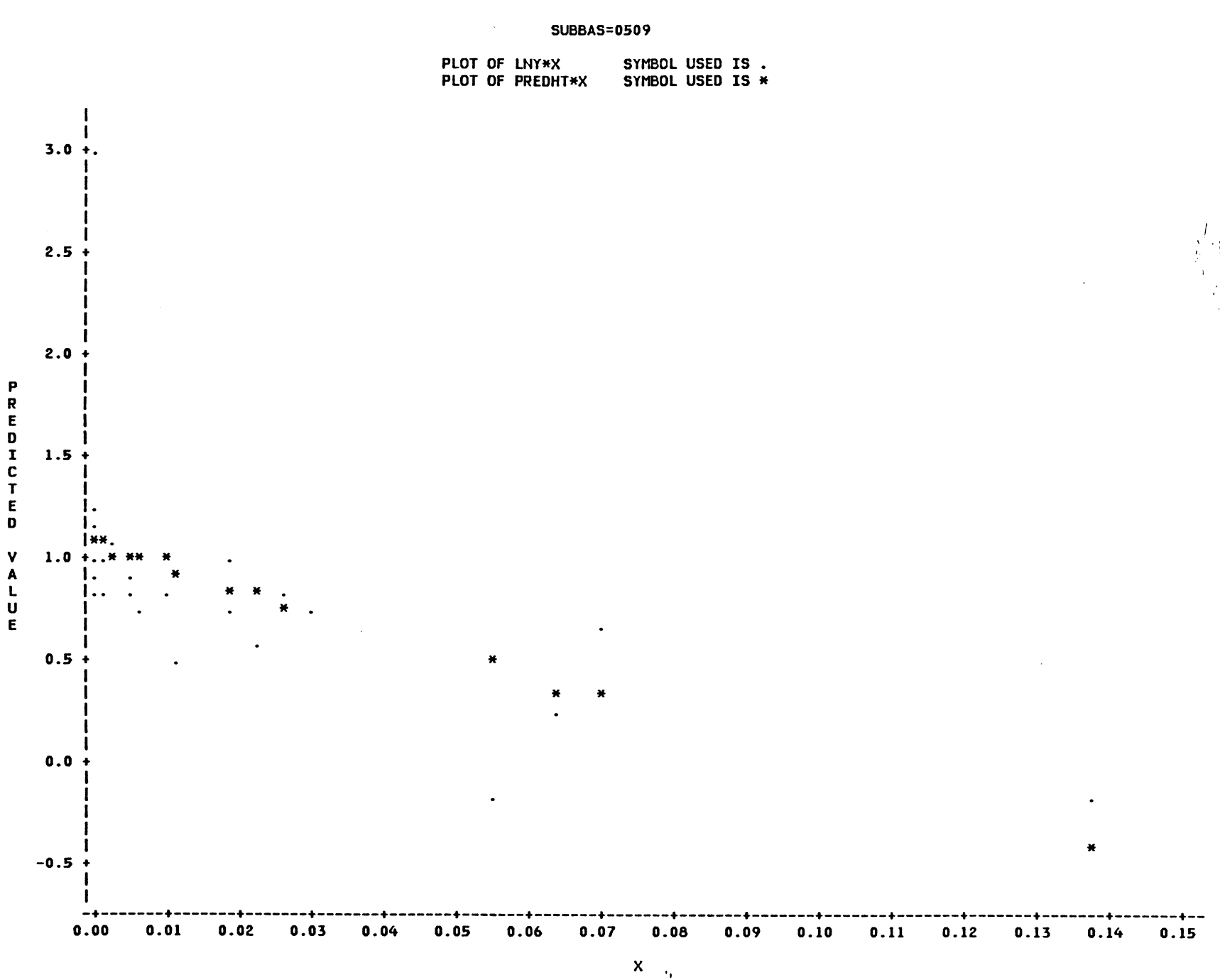

Figure 4-7. Actual data and predicted exponential equation for Subbasin 0509 with one outlier deleted (Pattern 2).

shows an excellent linear fit. Figure 4-7 represents the data for subbasin 0509, Pattern 2, and also shows an excellent linear fit with the deletion of one outlier.

4.2.3 Results of Statistical Regression Analysis for Each Subbasin

The regression least squares estimation procedure was used to fit an exponential regression model for each of the subbasin data sets. The regression estimation results are presented in Appendix C by subbasin for each basin. The results in Appendix C were obtained by the procedures described in the following paragraphs.

The transformed data, ln(y), for each subbasin were examined for outliers. An outlying data point is that point with a ln(y) value that exceeds three standard deviations from the mean of the ln(y) values at a given x value. Such data points are not considered to be representative of the population from which the sample is drawn (i.e., the hydrologic conditions of that subbasin) and, as such, the outliers were deleted. The regression statistical analysis of ln(y) on x assumes that the conditional distribution of ln(y), for a given value of x, is normal; therefore, approximately 99.7 percent of the ln(y) data, at a given value of x, should fall within ± 3 standard deviations from the mean at that value of x. Outliers were detected in data for subbasins that belonged to Pattern 2, as described in Section 4.2.2 (Figure 4-2), and these outliers were deleted.

After data were examined for the presence of outliers, a linear fit was examined for the regression of ln(y) on x in each subbasin data set. The significance of the correlation coefficient was tested in each subbasin data set. The linear regression of ln(y) on x, and hence the exponential regression of y on x, was used for each subbasin with a significant correlation coefficient. A correlation coefficient is considered significantly different from zero if the p-value (probability of exceeding the corresponding t-value) is less than the significance level 0.05. If the p-value is less than 0.05, there is at least

95 percent probability that the correlation coefficient is not equal to zero. Each subbasin that belonged to Pattern 2 (Section 4.2.2) showed significant correlations between ln(y) and x following deletion of the outlier(s).

The nonsignificance of the correlation coefficients between ln(y) and x for some subbasin data sets implies that the regression slope of ln(y) on x is not significantly different from zero. Therefore, the best prediction model for ln(y) in these cases is ln(y) = constant, where the constant is the mean of ln(y). Subbasins that belonged to Pattern 3 or 4 (Section 4.2.2) showed nonsignificant correlations and hence were modeled by the constant model.

The data for subbasins that belonged to Pattern 5 (Section 4.2.2) (i.e., few data points) were combined with the data of other subbasins. Geographical locations (established by merging data for a subbasin with adjacent subbasins) and expert judgment were used in deciding which subbasins would be combined. The combined data were examined for outliers and significant correlations. Combined data with significant correlations were fitted by the linear regression model of ln(y) on x ($y = e^{A-Bx}$). Combined data with nonsignificant correlation were fitted by the constant model (ln(y) = constant). The combining of data sets of subbasins that belonged to Pattern 5 did not alter the regression characteristics of the adjacent subbasin data set.

In Versar (1984), the regression of y on x was also modeled by the exponential model for each of Basins 1 through 7, 9 through 12, and 14, 15, 17, and 18. The regression equations in Basins 8, 13, and 16 were fitted by the constant model. These results are presented in Appendix C for comparison.

A comparison of the regression results in Appendix C shows the following:

- In most of the subbasins, there is a significant correlation between y and x, and the "best" regression model is

$$y = e^{A-Bx}.$$

- For most of the subbasins, the correlation between ln(y) and x for a specific subbasin exceeded the correlation between ln(y) and x for the combined data set for the entire basin.

- The regression results varied significantly among the subbasins of each basin.

- All the B values are positive for those subbasin data sets having significant correlations. This indicates a negative slope for the regression line of ln(y) on x.

4.3 Incorporation of Coefficients into PDM3

The old (1986) EED Basic PDM program used coefficients of variation for basin flows. For PDM3, as described above, coefficients for subbasin flows were developed and placed into the Pascal program coding. Thus, when a user selects a particular reach, the coefficient for the subbasin the reach belongs is used in the PDM calculation. The original basin coefficients were also placed into the PDM3 coding. This was done in case a user wants to perform a PDM run for no particular subbasin in a basin. Finally, from the old EED version, a coefficient of variation for flow was developed for all the basins combined. This coefficient was developed in case the user does not specify any basin. This coefficient was placed in PDM3 as well.

5. OPTION 1 - REACHES WITHOUT GAGING STATIONS - DERIVATION OF PROBABILITY CALCULATIONS AND EXCEEDANCE CALCULATIONS

5.1 Mathematical Derivation of the Probability of Exceedance

The probability dilution model (PDM) presented by Di Toro (1984) and EPA (1984) is described by the equation

$$CT = \frac{QS}{QS + QE} CS + \frac{QE}{QS + QE} CE \qquad (1)$$

where

CT = the downstream concentration
QS = the upstream flow
QE = the effluent discharge flow
CS = the upstream concentration
CE = the effluent concentration.

The following simplified form of the model is obtained by assuming that the upstream concentration CS = 0.

$$CT = \frac{QE}{QS + QE} CE \qquad (2)$$

Di Toro (1984) presented a direct evaluation method and an approximate moment method to compute the probability of CT exceeding a certain value of concern CT* using the general form (1) of the PDM and assuming that the four input variables of the model CS, QS, CE and QE, are jointly (dependent) lognormally distributed. The computer source code used to compute the probability numerically uses the special form (2) of the PDM, assuming that the three input variables QS, CE and QE are independent and each has a lognormal distribution. This assumption of independence is made because there is no data to estimate the correlations among QS, QE, and CE.

The first part of this analysis is a step-by-step check of the match between the mathematical derivation of the exceedence probability (based on the simplified model (2)) as well as its match with the original EED computer source code (Appendix D). This is performed by showing the

mathematical derivation of the exceedence probability using the simplified model (2) followed by one numerical presentation example of the computer source code.

5.1.1 Defining Probability in Terms of an Integral

The probability of CT exceeding a certain value CT* can be expressed as an integral of the joint probability density over the values of flows QS and QE and concentration CE for which CT > CT*. This requires an integral for each of the variables QS, QE and CE. One integral can be eliminated by using the following form of model (2):

$$CT = \frac{1}{1 + R} CE \tag{3}$$

where

$$R = QS/QE \tag{4}$$

defines a dilution ratio.

Therefore,

$$Pr(CT > CT^*) = Pr\left[\frac{CE}{1 + R} > CT^*\right] \tag{5}$$

The limits, over the values of the dilution ratio R and the concentration CE, of the double integral of the joint probability density of R and CE are obtained as follows:

For the value CT*, the equality

$$CT^* = \frac{CE}{1 + R} \tag{6}$$

defines a surface in CE and R space and the required probability is the integral of the joint probability function above this surface. For any fixed R, $0<R<\infty$, equality (6) defines a point on the straight line

$$CE = (1 + R) \bullet CT^* \tag{7}$$

the region in which CT > CT* is the interval $CE = [(1 + R)\bullet CT^*, \infty]$ for $0<R<\infty$. Thus, the limits of the integrals for the exceedence probability are

$$Pr(CT > CT^*) = \int_{R=0}^{\infty} \int_{CE1}^{\infty} f(CE,R)\, dCE\, dR \qquad (8)$$

in which

$$CE1 = (1 + R) \bullet CT^* \qquad (9)$$

and f(CE,R) is the joint probability density function for CE and R.

5.1.2 Simplifying the Integral Limits

The dilution ratio (R = QS/QE) has a lognormal distribution when both stream flow (QS) and effluent flow (QE) are lognormal. The dilution ratio (R) and effluent concentration (CE) are statistically independent since it has been assumed that the variables QS, QE and CE are pairwise statistically independent. Therefore, the joint probability density function f(CE,R) can be expressed as the product of the probability density functions f(CE) and f(R), i.e.

$$f(CE,R) = f(CE)\, f(R). \qquad (10)$$

Substituting (10) in (8), gives

$$Pr(CT > CT^*) = \int_{R=0}^{\infty} f(R) \left[\int_{CE1}^{\infty} f(CE)\, dCE \right] dR. \qquad (11)$$

The following inverse probability transformation removes the probability density f(R) from the integral and results in a finite range of integration. Using the variable of integration

$$X = \int_{0}^{R} f(R)\, dR \qquad (12)$$

then

$$dX = f(R)\, dR$$

so that the exceedence probability (11) becomes

$$Pr(CT > CT^*) = \int_{X=0}^{1} \left[\int_{CE2}^{\infty} f(CE)\, dCE \right] dX \qquad (13)$$

in which

$$CE2 = (1 + D(X)) \bullet CT^*; \qquad (14)$$

D(X) is derived below.

5.1.3 Characterize the Probability Distribution of CE and R

The probability distribution of the input variables QS, QE and CE in the PDM (2) are characterized by their means (μ) and coefficients of variation (V). The following notations of the means and coefficients of variation are used throughout the mathematical derivation and the computer source code.

Variable	Mean		Coefficient of Variation	
	Mathematical Derviation	Computer Source Code	Mathematical Derivation	Computer Source Code
Stream flow (QS)	μ(QS)	MQS	V(QS)	V1
Effluent flow (QE)	μ(QE)	QE	V(QE)	V2
Effluent concentration (CE)	μ(CE)	CE	V(CE)	V3

5.1.4 Characterization of the Log Transformation of CE and R

The exceedence probability (13) is more conveniently expressed in terms of the (natural) logs of the variables defined by

$$ce = \ln(CE)$$

and

$$r = \ln(R).$$

The transformed variables ce and r are normally distributed when the variables CE and R are lognormally distributed. The probability

distributions of ce and r are characterized by their means (μ) and standard deviations (σ). The following notations were used to denote the means and standard deviations of ce and r in the mathematical derivation and the computer source code.

Variable	Mean		Standard Deviation	
	Mathematical Derviation	Computer Source Code	Mathematical Derivation	Computer Source Code
Log of effluent concentration (ce)	μ(ce)	U3	σ(ce)	W3
Log of dilution ratio (r)	μ(r)	U9	σ(r)	W9

The means and standard deviations of ce and r are derived from the means and coefficients of variations of QS, QE and CE as follows:

$$\mu(ce) = \ln\left[\mu(CE)/\sqrt{1 + V2(CE)}\right] \quad (15)$$

$$\sigma(ce) = \sqrt{\ln[1 + V2(CE)]} \quad (16)$$

$$\mu(r) = \ln\left[\mu(QS)/\mu(QE)\right] + \ln\sqrt{1 + V2(QE)} - \ln\sqrt{[1 + V2(QS)]}. \quad (17)$$

$$\sigma(r) = \sqrt{\ln[1 + V2(QS)] + \ln\left[1 + V2(QE)\right]} \quad (18)$$

5.1.5 Correction of μ(CE) and μ(r)

A normalization scheme was introduced in the Technical Guidance Manual for Performing Waste Load Allocations Book VII, prepared for the Office of Water Regulations and Standards (1984). This normalization scheme was used in the computer source code of the PDM. The following are the assumptions used in the scheme and their consequences on the characterization parameters of the variables ce and r.

The stream target concentration CT is produced when the discharge flow is the mean effluent flow μ(QE) and the stream flow is equal to the design value (low flow (7Q10)). Therefore, the input value QE of effluent flow is used as the mean effluent flow and the following corrections were introduced.

$$\text{Corrected Mean Stream Flow (CMQS)} = \text{Mean Stream Flow} - \text{Effluent Flow (QE)} \quad (19)$$

If Low Stream Flow > Effluent Flow:

$$\text{Corrected Low Stream Flow (CQS7Q10)} = \text{Low Stream Flow (7Q10)} - \text{Effluent Flow (QE)} \quad (20a)$$

If Low Stream Flow $\leq$ Effluent Flow:

$$\text{Corrected Low Stream Flow (CQS7Q10)} = \text{Low Stream Flow (7Q10)}. \quad (20b)$$

Using the following ratios,

$$F1 = CQS7Q10/CMQS \quad (21)$$

$$F2 = CQS7Q10/QE \quad (22)$$

the means of ce and r in (15) and (17) are replaced by (23) and (24) below, respectively.

$$\mu(ce) = \ln \ (1 + F2)/\sqrt{1 + V2(CE)} \quad (23)$$

$$\mu(r) = \ln(F2/F1) + \ln\sqrt{1 + V2(QE)} - \ln\sqrt{1 + V2(QS)} \quad (24)$$

5.1.6 Correction of User Input of Concentration of Concern

The user input value of the concentration of concern (denoted by PEP in the computer source code) is normalized as follows:

The concentration of chemical in the effluent (CE) is computed as:

$$CE = \text{Amount of Chemical Released/Effluent Flow} = REL/QE. \quad (25)$$

The stream concern concentration level (CL) is given by

$$CL = (QE * CE)/(QE + CQS7Q10). \quad (26)$$

The normalized concentration of concern is

$$CT^* = PEP/CL. \tag{27}$$

Relationships (25), (26), and (27) show that the normalized concentration of concern is inversely related to the amount of chemical released (REL) as follows:

$$CT^* = (PEP/REL)\bullet(QE + CQS7Q10). \tag{28}$$

5.1.7 Computation of the Inside Integral of Equation 13

The inside integral in (13) is evaluated using the transformed variable ce as follows:

$$\int_{CE2}^{\infty} f(CE)dCE = Pr(CE > CE2)$$
$$= Pr\ (ce > \ln(CE2)). \tag{29}$$

Since the variable CE has a lognormal distribution, the transformed variable ce has a normal distribution with a mean and standard deviation defined by (23) and (16), respectively. The probability in (29) can be expressed in terms of the cumulative probability function Q(Z) of a standard normal random variable Z;

$$Pr\ (ce > \ln\ (CE2)) = Q\left[\frac{\ln(CE2) - \mu(ce)}{\sigma(ce)}\right]. \tag{30}$$

From (14),

$$\ln(CE2) = \ln\ (CT^*) + \ln\ (1 + D(X)).$$

Therefore,

$$\int_{CE2}^{\infty} f(CE)\ dCE = Q\left[\frac{\ln CT^* + \ln\ (1 + D(X)) - \mu(ce)}{\sigma(ce)}\right]. \tag{31}$$

The inverse function D(X) is computed from the inverse of the cumulative distribution function:

$$P(Z^*) = Pr(Z < Z^*) = 1 - Q(Z^*).$$

That is, if $p = P(Z^*)$, then the inverse function, P^{-1} satisfies the relationship

$$Z^* = P^{-1}(p).$$

Thus, from (12)

$$\begin{aligned} X &= Pr\ (R < D(X)) \\ &= Pr\ (r < \ln D(X)) \\ &= Pr\left[Z < \frac{\ln D(X) - \mu(r)}{\sigma(r)}\right] \\ &= Pr\left[\frac{\ln D(X) - \mu(r)}{\sigma(r)}\right] \end{aligned}$$

which leads to

$$D(X) = \exp\ (\mu(r) + \sigma(r)\ P^{-1}(X)). \tag{32}$$

Let

$$G(X) = \ln(1 + D(X)) - \mu(ce) \tag{33}$$

Therefore,

$$\int_{CE2}^{\infty} f(CE)\ dCE = Q\left[\frac{\ln CT^* + G(X)}{\sigma(CE)}\right]. \tag{34}$$

5.1.8 Simplifying the Double Integral Equation 13

Substituting for the inside integral of (13) from (34) yields the following expression for the exceedence probability:

$$Pr(CT > CT^*) = \int_{X=0}^{1} Q(X)\ dX \tag{35}$$

where Q(X) is defined in (34).

5.1.9 Quadrature Method for Computing the Integral in Equation 35

The finite integral defined in (35) is evaluated numerically using the quadrature method. The quadrature method divides the distance on the

X-axis between 0 and 1 into 32 unequal subdistances denoted by a_i, i = 1, ..., 32 (denoted by Z5(i), i = 1, ..., 32 in the computer source code). These distances represent the width of the region (U_{i-1}, U_i), i = 1, ..., 32, where $U_0 = 0$ and $U_{32} = 1$. The method also defines 32 points (X_i, i=1, ...32, in the regions (U_{i-1}, U_i), (one each) and computes the value of Q(X) at each X_i, i = 1, ..., 32 using (31), (33), and (34). The values of X_i are denoted by R5(i), i = 1, ..., 32 in the computer source code. The numerical value of (33) is then computed as

$$Pr(CT > CT^*) = \sum_{i=1}^{32} a_i \, Q(X_i). \qquad (36)$$

5.2 Computation Procedure of Probability of Exceedance

The procedure of computing the exceedance probability in (36) is the following:

(1) Find the 32 pairs of constants a_i and X_i (i = 1, ..., 32) as defined by the quadrature method. These constants are called the Laguerre roots and weights and are computed in two computer subroutines in the program. (Source Code 3800 to 4740, Appendix G).

(2) Find 32 values of the inverse function $P^{-1}(X)$ at the X_i values using function 26•2•23 of Abramowitz and Stegun (1954). The procedure for computing $P^{-1}(X)$ is given in the subroutine defined by computer source code 3640 to 3850. Six constants E1, ..., E6 are used in the procedure and are defined in lines 3500 to 3550 (Appendix G). The values of a_i, X_i and $P^{-1}(X_i)$ are presented in Table 5-1.

(3) Use the user's and default value for the parameters of the input variables Qx, QE, and CE and the normalization procedure (Technical Guidance Manual for Performing Waste Load Allocation Book VII) to find the means and standard deviations $\mu(r)$, $\sigma(r)$, $\mu(Ce)$, and $\sigma(Ce)$ of the log transformed

Table 5-1. The Quadrature Method Constants (a_i, X_i) and the Inverse Function $P^{-1}(X_i)$

i	(1) a_i	(2) X_i	(3) $P^{-1}(X_i)$
1	.006788114855	5.299532500000002E-03	2.55605685
2	.015563380985	.0277124885	1.91595462
3	.02378962792	.0671843988	1.49739139
4	.031157242825	.1222977958	1.16367273
5	.0373989972	.1910618778	.87385280
6	.04228912985	.2709916112	.60946955
7	.04565085385	.3591982246	.36015847
8	.047362652625	.45249374508	.11909308
9	.006788114855	.9947004675	-2.5560569
10	.015563380985	.9722875115	-1.9159546
11	.02378962792	.9328156012	-1.4973914
12	.031157242825	.8777022042	-1.1636727
13	.0373989972	.8089381222	-.87385280
14	.04228912985	.7290083888	-.60946955
15	.04565085385	.6408017754	-.36015847
16	.047362652625	.54750625492	-.11909308
17	2.070731185E-22	3.51942972235154E-23	8.75687846
18	2.52523685E-18	6.102293358538447E-19	8.75687846
19	3.1489835015E-15	9.563603367893601E-16	7.94648410
20	1.0635395165E-12	3.876279993601101E-13	7.16502049
21	1.43117512E-10	6.125926849120944E-11	6.43577147
22	9.405124205E-09	4.679119314232594E-11	5.74162921
23	3.4141596655WE-07	4.679119314232594E-11	5.07188075
24	.000007422293435	4.959413672622759E-07	4.41884934
25	.00010213595765	7.96144813648425E-05	3.77630691
26	.000924535471765	8.499454124830663E-04	3.13847642
27	.00564995004015	6.232245825231075E-03	2.49913254
28	.02366446434705	3.215824067592621E-02	1.85039317
29	.068148467146	.1189224496484803	1.18049873
30	.132897888822	.3194809257984161	.46872744
31	.165532892274755	.6295837759971619	-.33030947
32	.103075857479	.9160820245742798	-1.3794268

(1) Denoted by Z5 in the computer source code.
(2) Denoted by R5 in the computer source code.
(3) Denoted by X9 in the computer source code.

variable r and ce. Equations (19), (20), (21), and (22) are used to compute the normalized values and equations (16), (18), (23) and (24) are used to compute the means and standard deviations.

The user input values are:

- Mean Stream Flow (MQS)
- 7Q10 Stream Flow (QS7Q10)
- Effluent Flow (QE)
- Amount of chemical released (RREL), kg/site/day
- The number of release days (DAY)

  The basin and sub-basin hydrological number (to find the coefficient of variation of stream flow using the basin/sub-basin regression equation).
- The coefficient of variation of effluent flow (default value is 0.24).
- The coefficient of variation of effluent concentration (default value is 0.85).

(4) Use the computed values of $\mu(r)$ and $\sigma(r)$ and the values of the inverse function $P^{-1}(X_i)$ (Table 5-1) to find the corresponding values of $D(X_i)$ defined in equation (31). Use the $D(X_i)$ value and $\mu(ce)$ to find the $G(X_i)$ values defined in equation (33).

(5) Use the normalization procedure in equations (25), (26), and (27) and the user input value of the concentration of concern to find the normalized value CT*. Use the values of CT*, $G(X_i)$, and $\sigma(ce)$ to find the argument of the $Q(\bullet)$ function.

(6) Use Abramowitz and Stegun (1954) procedure to find the value of Q(X) at the value of the argument computed in step (5) above. Q(Z) is the area under the standard normal density function to the right of Z and can be found in most of the statistics text books. Abramowitz and Stegun (1954) procedure is given in the computer source code lines 2580 to 2670 (Appendix D).

(7) Use the a_i values defined in Table 5-1 and values of the Q(X) function to find the exceedance probability according to equation (36).

5.3 Example Computation

The two previous subsections described the calculation of the probability of exceedance. This subsection describes its use in the final PDM3 Option 1 output for reaches without gaging stations. The use of each of the user inputs for Option 1 is demonstrated as well.

User input values:

- Mean Stream Flow = 1000 mld.
- 7Q10 Stream Flow = 100 mld.
- Effluent Flow = 10 mld.
- Amount of chemical released = 10 kg/site/day.
- The number of release days = 250 days.
- The basin number = 03.
- The sub-basin number = 02.
- The coefficient of variation of effluent flow = default (0.24).
- The coefficient of variation of effluent concentration = default (0.85)
- The concentration of concern = 0.5 ug/l

Coefficient of variation of stream flow: using the regression equation for basin 03 sub-basin 02 the coefficient of variation of stream flow is:

$$V1 = \exp(0.759 - 2.265X)$$
$$= 1.703185$$

where

$$X = \text{7Q10 Stream Flow/Mean Stream Flow}$$
$$= 0.10$$

The normalization constants F1 and F2 are as follows:

F1 = Corrected Low Stream Flow/Corrected Mean Stream Flow

= (100 - 10)/(1000 - 10) = 0.090909

F2 = Corrected Low Stream Flow/Effluent Flow

= (100 - 10)/10 = 9

The Mean and Standard Deviation of ce: using equations (23) and (16) the following values are obtained:

$$\mu(ce) = 2.03069673$$

$$\sigma(ce) = 0.73741218.$$

The Mean and Standard Deviation of r: using equations (24) and (17) the following values are obtained:

$$\mu(r) = 3.94252512$$

$$\sigma(r) = 1.19045951.$$

D(X) and G(X): using the values of the inverse function $P^{-1}(X)$ defined in Table 5-1 and the values of $\mu(r)$ and $\sigma(r)$ the values of D(X) are computed according to equation (31); using the values of D(X) and $\mu(ce)$ the values of G(X) are computed according to equation (33). The values of G(X) are given in Table 5-2. The values of G(X) are denoted by Z9 in the computer source code.

The Normalized Concentration of Concern: using equations (25), (26) and (27) the normalized value CT* is obtained;

$$CE = 10000/10 = 1000 \ \mu g/l$$

$$CL = (10)(1000)/(10 + 90) = 100 \ \mu g/l$$

$$CT^* = 0.5/100 = 0.005 \ \mu g/l$$

Q(X): using the normalized value CT* and the values of G(X) given in Table 5-2, the values of Q(X) are obtained using Abramowitz and Stegun (1954) procedure defined in the source code lines 2580 to 2670 (Appendix G). The abbreviated values of Q(X) are given in Table 5-2.

The probability of exceedance: using the constants a_i defined in Table 5-1 and the Q(X) values in Table 5-2, the probability of exceedance is computed from equation (36) and is given by

Table 5-2. G(x) and Q(x) Values of the Quadrature Method for the Input Values of the Example Computation

i	$G(x)^{(1)}$	$Q(x)^{(2)}$
1	4.956	.6789
2	4.194	.9327
3	3.697	.9850
4	3.301	.9966
5	2.958	.9992
6	2.646	.9998
7	2.353	.9999
8	2.070	.9999
9	-.7897	.9999
10	-.1952	.9999
11	.2383	.9999
12	.6011	.9999
13	.9249	.9999
14	1.2255	.9999
15	1.5124	.9999
16	1.7921	.9999
17	12.336	.0000
18	12.336	.0000
19	11.371	.0000
20	10.441	.0000
21	9.5733	.0000
22	8.7470	.0000
23	7.9497	.0001
24	7.1723	.0055
25	6.4075	.0662
26	5.6485	.3174
27	4.8879	.7110
28	4.1167	.9454
29	3.3219	.9963
30	2.4808	.9999
31	1.5469	.9999
32	.3651	.9999

(1) Denoted by Z9 in the computer source code.
(2) Denoted by F in the computer source code.

$$P\ (CT\text{-}CT^*) = 0.99533.$$

This probability is then used to compute the number of days per year where the concentration of concern CT* is exceeded as follows:

Days per year concentration CT* is exceeded
= (probability of exceedance) (number of release days)
= (0.99533) (250)
= 248.8324 days.

6. DOCUMENTATION FOR OPTION 2 - WORST CASE ANALYSIS FOR A FACILITY IN A SPECIFIC INDUSTRIAL CATEGORY

The second of the two PDM3 options is for non-site-specific scenarios. In this option, the user needs only to know the SIC code to which the facility of concern belongs (in addition to concentration of concern, loading, and release days) in order to run the model. This option uses the same PDM calculation as Option 1 except that it uses it multiple times to calculate a worst case probability of exceedance for a SIC group. This analysis is used to represent the 10 percent of facilities for the SIC group that yield the highest exceedance probabilities.

EED had a version of this option (Dilute) and there were four main objectives to improving it:

1. Combine data for direct and indirect dischargers for each of the 39 SIC groups and use them in the probability calculations;

2. Write the program so that all facility plant flows and receiving stream flows are used in the probability calculation and then determine the 10 percentile probability;

3. Incorporate the appropriate sub-basin coefficient of variation into each stream flow when calculating probability; and

4. Reduce the running time of the program so that it is feasible to use.

The first three objectives could be achieved by (a) developing detailed programs to access, analyze, and summarize data not used previously and (b) adding additional steps to the probability calculations. The fourth objective, however, could not be accomplished by simply altering the PDM programming. Since the probability calculations require time-consuming steps that far exceed the ability of a PC to perform quickly, it was necessary to develop another method to obtain quick results. The method chosen was essentially to run the calculations hundreds of times for each facility in each SIC group on the IBM mainframe computer and to store the probabilities of exceedance on PC diskette files. The

probability for a given release rate and concern level could be obtained from these files by interpolation.

This section describes the programs, calculations, and analyses used to obtain the above-listed objectives. The section is arranged in the order in which the work was performed. Figure 6-1 is a flow diagram depicting the phases of the work performed and the relative order in which they were performed.

6.1 Retrieval and Arrangement of Data from HLDF

The initial stage of this task was to retrieve the data and arrange them so that they could be used in the program. Data from the HLDF files were obtained using specified data templates and JCL. SAS® procedures were used to write a program to merge, analyze, and arrange the retrieved HLDF data. Although there are several distinct steps to this phase, they are actually carried out by one program that has several smaller programs in it (Section 6.1.3). The programmer's documentation prepared under a separate cover contains copies of the program used to accomplish this work and an explanation of each of the program's functions.

6.1.1 IFD Retrievals

The IFD file contains information on both direct and indirect dischargers and the object here was to combine these data by SIC. The information on direct dischargers includes their name, NPDES number, SIC code, effluent flow, the type of effluent (processing; cooling), the discharge code (surface, municipal), and the receiving stream. Indirect dischargers are listed according to the POTW that receives their effluent. Other information on indirect dischargers includes their SIC code, effluent flow to the POTW, and the receiving stream of the POTW.

The retrievals from IFD were performed by SIC code of a facility's effluent pipes. That is, a template was written specifying that any facility with a pipe of a certain SIC code be retrieved along with its pertinent data. Because many plants have multiple pipes with different processes, each pipe is assigned a SIC code. In many cases, the pipe SIC

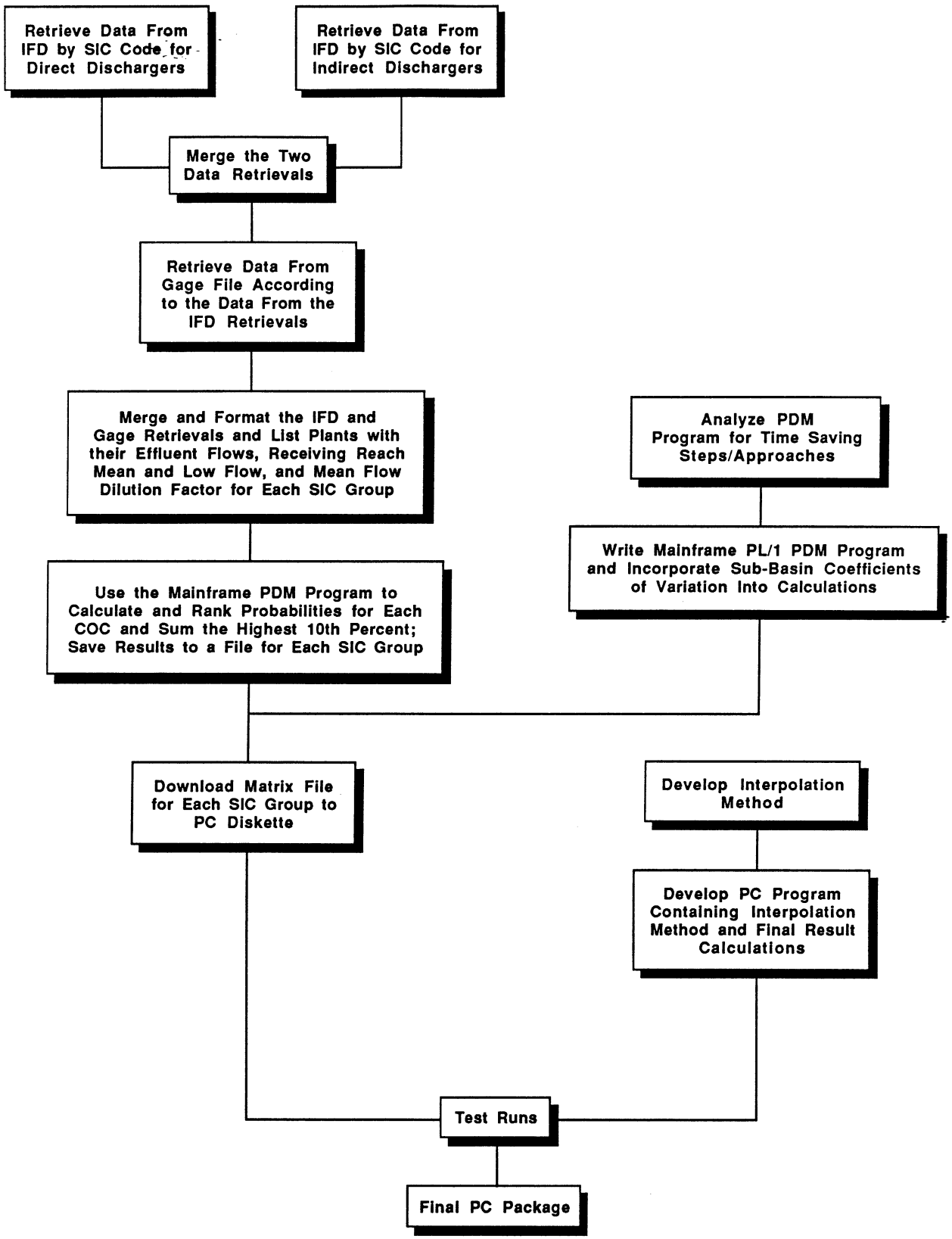

Figure 6-1. Flow Diagram for Development of Option 2.

code will differ from the facility's overall SIC code. Thus, the retrievals were performed according to pipe SIC codes. Table 2-2 lists the SIC codes for which data were retrieved.

Because of the makeup of the IFD file, it is not possible to retrieve both direct and indirect dischargers in the same retrieval in the form required for this task. For this reason, it was necessary to run two retrievals for each SIC group, one each for directs and indirects. Finally, only those facilities having processing effluents were retrieved; facilities that only discharge cooling water were omitted from the retrievals.

The following data elements were retrieved and saved for direct dischargers:

- NPDES number
- Facility name
- Number of pipes
- Effluent flow for each pipe
- Discharge code for each pipe
- Receiving stream reach number
- Hit code

The following data elements were retrieved and saved for indirect dischargers:

- Facility name
- NPDES number of receiving POTW
- Effluent flow of POTW
- The POTW's receiving stream number
- Hit code.

If an indirect discharger has multiple pipes to a POTW, it and the POTW are listed only once; the POTW's effluent flow is not increased to reflect the number of pipes.

It is important to note that not all of the faclities retrieved could be used in the PDM calculations. Only those with a complete set of flow data (effluent flow, reach mean and low flows) could be used. For many utilities, the effluent flow is missing from IFD. The reach flows could

be missing for three reasons: (1) receiving reach number not identified in IFD, (2) reach could be a water body without any flow data (shoreline of a bay, lake, estuary, wide river, or ocean), (3) the flow listed for the reach is bad (this is indicated as such in the GAGE File). Table 6-1 lists the total number of facilities with complete flow data for each SIC group.

The retrievals were saved as data set files and a program using SAS® procedures was written to merge, edit, and arrange them (see Section 6.1.3 below).

6.1.2 GAGE Retrieval

Among the data retrieved for the above facilities from IFD was the reach number of the receiving stream (for indirects, the receiving stream of the POTW). The reach numbers were used to retrieve flow data from the GAGE file. This was accomplished by writing a program that would read the reach numbers from the direct and indirect files created from the IFD retrievals. After reading these data files, the program accessed the GAGE file for each reach and obtained the estimated mean and low flow values; the flow data were then saved in a separate file. Estimated flows instead of USGS flows were used primarily because estimated flows are always for the downstream end of each reach whereas gaging stations may be located anywhere on a reach.

The files created in this step were then merged and formatted with the IFD information (see 6.1.3).

6.1.3 The COMBINE Program

The COMBINE program was written to access the files created from the IFD and the GAGE retrievals and to merge, edit, and format them together. The programs written for each of the IFD retrievals and the GAGE retrieval were incorporated into the COMBINE program such that all the work necessary for a single SIC group could be done in one shot. The program uses SAS® procedures to accomplish most of the merging, editing, and formatting.

Table 6-1. Summation of SIC Group Data Files

SIC Codes	Industry	Number of facilities with complete flow data[1]
2891	Adhesives and Sealants Manufacture	24
3674, 3679	Electronic Components Manufacture	147
3411	Metal Can Manufacture	35
3471	Electroplating	1,858
332, 336	Foundries	255
2865, 2869	Organic Chemicals Manufacture	252
2893	Ink Formulation	11
281	Inorganic Chemicals Manufacture	348
3111	Leather Tanning & Finishing	83
2911, 2992	Lubricant Manufacturers	174
(See * Below)	Metal Finishing	1,244
2851	Paint Formulation	57
101-109	Ore Mining & Dressing	219
2621, 2631, 2661	Paper and Paperboard Mills	407
2621	Paper Mills, except Building Paper Mills	241
2631	Paperboard Mills	130
2661	Building Paper and Board Mills	37
2819, 2869, 2879	Pesticides Manufacture	424
2911	Petroleum Refining	156
7221, 7333, 7395, 7819	Photographic Processing	41
3079	Plastic Products Manufacture	223
2821, 2823, 2824	Plastic Resins and Synthetic Fibers Manuf.	229
(See ** Below)	POTWs (Industrial)	1,036
4952	POTWs (All)	12,646
271-277	Printing	157
2611	Pulp Mills	70
3011, 3021, 3031, 3041	Rubber Products Manufacture	100
2841, 2842, 2843, 2844	Soaps, Detergents, etc. Manufacture	108
7211, 7213-7219, 7542	Auto and other Laundries	535
3711, 3713	Motor Vehicle Manufacture	70
3631, 3632, 3633, 3639, 3431, 3469	Large Household Appliance and Parts Manufacture	137
3315-3317, 3351-3357, 3463, 3497	Primary Metal Forming	281
2281, 2282, 2283, 2284	Yarn and Thread Mills	128
2271, 2272, 2279	Carpet Dyeing and Finishing	99
2231	Wool Dyeing and Finishing	65
225, 2292	Knit Fabric Dyeing and Finishing	166
2261, 2262, 2269	Woven Fabric Dyeing and Finishing	167
2231, 225, 226, 2292	All Textile Dyeing, except Carpets	398
4911	Steam Electric Plants	536

* Metal Finishing = 3411-3462, 3465-3471, 3482-3599, 3613-3623, 3629, 3634-3636, 3643-3651, 3661-3671, 3673, 3676-3678, 3693-3694, 3699, 3711-3841, 3851, 3873-3999

** POTWs (Industrialized) = IDSI = 1011-1999, 2211-5199, 5511-5599, 7211-8099

[1] Plant flow and receiving stream mean and low flows available.

In addition to merging and formatting the data, the COMBINE program also performs the following calculations:

- Calculates total effluent pipes and flows for facilities with multiple pipes.
- Converts effluent pipe flows from million gallons per day to million liters per day.
- Converts reach flows from cubic feet per second to million liters per day.
- Substitutes plant effluent flow for reach mean flow when it exceeds the reach mean flow; does the same for the reach low flow.
- Calculates the mean flow dilution factor (reach mean flow ÷ effluent flow); if the effluent flow exceeds the mean flow, the dilution factor is set equal to one.
- Sums the number of facilities retrieved for each SIC group and counts the number with complete flow data (see Table 6-1).

The first output file produced was reviewed to make certain all the procedures ran correctly. The subsequent files were checked to make certain that the correct facilities were specified in the retrievals and that the programs ran to completion. The files are stored together on a tape at NCC in Research Triangle Park, NC.

6.2 Analyses of Probability Calculations for Time Saving Steps

The programming steps for the probability calculations were analyzed to see if they could be made to run faster. The first attempt to shorten the program involved its compilation. This action reduced the running time by 50 percent but such a reduction was not nearly enough to make the running time user friendly.

A second attempt involved examining the calculation steps of the program. The program has two main calculation loops. The first loop involves the inverse probability transformation, while the second involves the quadrature loop of the calculation. The first loop could be made to perform prior to running the program (as an initialization step) and

would not need to be recalculated again since it is independent of user input. This action would reduce the running time by as much as one-third. The running time, however, would still be too long considering the size of many of the SIC group files. The second loop is dependent on user input, specifically, the concentration of concern and the amount of chemical released. Because of this dependence, it was not possible to modify the loop to improve running time. It was decided that the running time of the program could not be significantly reduced by adjusting the programming steps. If the running time were to be reduced, an alternate approach to the program would be required.

It was decided that the only feasible approach to reducing the running time would be to create matrix files of probabilities and place them on PC diskettes. Probabilities would be determined for predefined values of the concentration of concern and the amount of chemical released and saved on diskette. These values could be accessed and reported by a PC very quickly. The calculations would be done on the EPA IBM mainframe computer. The only problem with this approach was that the combinations of concentration of concern levels and chemical loadings could run into the millions. Even with the mainframe computer, the number of iterations would take too long to be reasonably performed. In addition, the resulting matrix files for each SIC group would be too large to store on PC diskettes.

The next step was to see if there was any way to reduce the number of concentration of concern levels to loading combinations. A close examination of the probability calculation revealed that there is a direct relationship between the concentration of concern and chemical loading. The relationship between the two is outlined below.

The probabilistic dilution model defines the probability of exceedance, PER, as the probability that the downstream concentration CO exceeds a certain value of concentration C_0:

$$\text{i.e.,} \quad PER = Pr\ (CO \geq C_0) \qquad (1)$$

The value of C_0 is normalized in terms of a stream target concentration (such as the chronic criteria concentration, CL), so that the calculation can be used for a wide variety of pollutants. Stream concentration is therefore expressed in terms of $\beta = C_0/CL$, β being a dimensionless unit of concentration. The parameters of stream concentration, CO, are also normalized using CL.

The stream target concentration, CL, is defined in the model as the ratio of the amount of chemical released by a plant, Rel, to a flow value (7Q10), LQ:

$$\text{i.e.,} \quad CL = \frac{REL}{LQ} \tag{2}$$

Therefore,

$$\beta = C_0 \; / \; (Rel/LQ) \tag{3}$$

$$\beta = (LQ) \; (C_0/Rel) \tag{4}$$

and the probability of exceedance is given by:

$$PER = Pr \; (\text{normalized } CO \geq \beta) \tag{5}$$

$$= Pr \; (\text{normalized } CO \geq LQ \; (C_0/Rel)) \tag{6}$$

Equation (6) shows that the probability, PER, will be the same for any fixed value of the ratio of C_0 to Rel at a given value of LQ. For example, if Rel = 2 and C_0 = 0.5, the PER = 0.25 and if Rel = 4 and C_0 = 1, the PER = 0.25.

Because of this relationship, it is not necessary to run the multitude of concentration of concern levels and loading combinations. The ratio between Rel and C_0 could be used to find the probability, PER, for only one value of Rel (say 1 kg/day) and several values of C_0; the results would be stored on diskette. The probabilities for different Rel values are the probabilities obtained at Rel = 1 kg/day and adjusted to the various levels of concentration.

To summarize, it is the ratio between the concentration of concern and chemical loading that is important; any given ratio will produce a

probability. Thus, the matrix files could be created by just varying the concentration of concern levels and keeping the loading constant. The file could be interpolated to estimate probabilities for ratios not represented in the file. The reading and interpolation of a matrix file could be achieved in a matter of seconds by a PC program.

This method was chosen as the means to reduce the running time of Option 2. Although the exact probability for many combinations of concentration of concern levels and loadings would not be calculated since they would be interpolated, the immense time savings would justify the slight difference in the interpolated probability value. Further, the difference would be far less than the variability in the accuracy of the original flow data.

6.3 Creation of Probability of Exceedance Matrix Files

The creation of the probability matrix files was the most important part of this work. It was divided into three separate functions: (1) development of a mainframe program that would access the data files previously created and perform the probability calculations, (2) determine the number and interval of concentration of concern levels to use in the mainframe program, and (3) the actual running of the program.

6.3.1 PDM Mainframe Program

The original PC program used by EED was employed as the guide to writing the program for the EPA IBM mainframe program. Prior to writing the mainframe program, the incorporation of the coefficients of variation for sub-basin flows were written into the PC program (see Section 4). This was done for two reasons: (1) to check that the coefficients would work in the program and (2) to have a PC program that could be used to check the results of the mainframe program. A copy of the modified PC program is found in Appendix D. The mainframe program is written in PL/1. This is the same program as the modified PC Basic program. A copy of the PL/1 program is found in the programmer's documentation.

The result of the PDM mainframe program is actually a sum of the highest 10th percentile probabilities for a particular concentration of concern level. That is, if there are 268 facilities in a particular SIC group, the program is run 268 times for that particular concern level using the flow data for each facility. The resulting 268 probabilities are then ranked and the top 27 are summed. The sum of the highest 10th percentile probabilities is then used in the final calculations (see Section 6.5). Thus, the analysis is worst case because only the highest probabilities are being summed and saved instead of all the probabilities.

In order to assure that the mainframe program was written correctly and was running properly it was tested against the modified PC program. Appendix I contains outputs from both the modified PC program and the mainframe program. As can be seen, the results from the two programs are the same for the same inputs. This assured the correctness of the mainframe program.

6.3.2 Concentration of Concern Levels

Before running the mainframe program, the number and interval of concentration of concern levels used to create that matrix file had to be determined. The range had to be large enough to account for a ratio of concern level to loading that a user would likely input. The number of values would have to be sufficient enough to yield close interpolation estimates but not so many that a lot of mainframe CPU time would be required or PC diskette storage capacity exceeded. The following range and interval of concern levels (μg/l) were used:

.00010	to	.00990	(by .00005)
.0100	to	.0990	(by .0005)
.100	to	.990	(by .005)
1.00	to	9.90	(by .05)
10.0	to	99.0	(by .5)
100	to	990	(by 5)
1,000	to	9,900	(by 50)
10,000	to	99,000	(by 500)

There are 1,450 values in these ranges and intervals. The range allows a user to input a concern level to loading ratio of no less than 0.0001 μg/l to 1 Kg/day (e.g., if a user inputs a loading of 10 Kg/day, the lowest concern level that could be analyzed is 0.0010 μg/l; 100 Kg/day, 0.010 μg/l). A ratio less than this would exceed the range of the matrix file and interpolation is not possible. On the other end of the scale, the user may input a very high concern level and a very low loading. Although this ratio will not be in the range of the matrix either, the probability of exceedance will only equal zero anyway.

6.3.3 Running of the Mainframe Program

After verifying the calculation loops of the mainframe program, additional steps had to be incorporated into the program's front end and back end. Instructions to access the SIC flow data files created by the program COMBINE were added to the front end. Instructions to save the calculated, ranked, and summed probabilities and arrange them into matrix files were added to the back end of the program.

When the program was first run, it took nearly an hour to complete a single SIC group. Because of the expense of CPU time, an effort was made to reduce the running time. Two modifications were made. The first had to do with the reading of the concern levels and did not affect the final result; it had only a slight effect on the running time. The second modification dealt with changing the precision of the program. The original PL/1 program had a precision of 24 places past the decimal point (double precision variables). This was changed to 6 places past the decimal point. The net result was a four-fold CPU time savings. The change, however, did affect the sum of probabilities but only very slightly (~ 0.001). When this value is divided by the number of facilties in an SIC group, it becomes insignificant in terms of the final result. This version of the program was used to create the matrix file for each SIC group.

6.4 Development of PC Option 2 Program

6.4.1 Matrix File Downloading

An error-free direct line connection to the EPA mainframe was used to download the matrix files from the mainframe to PC diskettes. Once on PC diskettes, the files were saved and identified according to their SIC group. To ensure complete transfer, the new PC files were compared to the files on the mainframe.

6.4.2 Interpolation Method

The next step in the program development was to determine the best way to interpolate the matrix files. When the user inputs one of the 1450 concern level/loading rate ratios that is represented in the matrix file, the corresponding sum of probabilities is accessed. Interpolation is necessary when the user inputs a concern level/loading rate ratio not represented in the file. The matrix file for SIC group 2865-2869 was used as a test file. It was analyzed for its sum of probabilities characteristics using preprogrammed SAS® plot and regression analysis programs.

A plot of the sum of the probabilities versus the concentration of concern/loading ratio used in the matrix file showed that the sum of probabilities is exponentially decreasing with increasing concentration of concern loading ratio. To further illustrate the exponential relationship the matrix file was divided into 8 segments. An analysis of each segment showed that the relationship between the sum of probabilities and concentration of concern/loading ratio is exponential of the form $y = \exp(a + bx)$, where y is the sum of probabilities and x is the concentration of concern/loading ratio and a and b are regression factors. The exponential regression of y on x is mathematically equivalent to the linear regression of the log transformed variable of the sum of probabilities, ln(y), on x, i.e., $y = \exp(a + bx)$ is equivalent to $\ln(y) = a + bx$. The coefficient of determination (R^2) for the regression of ln(y) on x in each segment was high (0.90 to 0.96). The analysis of the data also

showed that when the file is divided into a higher number of segments (more than 8) the R^2 values increased. This implies that the relationship between the sum of probabilities and the concentration of concern/loading ratio can be described as a piece-wise exponential, i.e., between every two consecutive values of concentration of concern/loading ratio in the matrix file, the sum of probabilities is exponentially decreasing. Therefore it was decided that the appropriate method for interpolating between each two consecutive values of concentration of concern/loading ratio is to use a piece-wise exponential interpolation.

The plan for interpolating the matrix files involved storing matrix files that were created on the mainframe computer on a PC diskette and use the piece-wise exponential interpolation method to obtain interpolated values when the concern level/loading ratio is not in the matrix file. The interpolation is conducted as follows:

1. The sum of probabilities for the immediate higher and lower ratios are accessed. The two concern level/loading ratio values selected from the matrix files and their corresponding sum of probabilities are used to find the coefficients a and b of the linear relation- ship (of the log transformed sum of probabilities with the concern level/loading ratio).

   $b = [\ln(SUM2) - \ln(SUM1)]/(COC1 - COC2)$
   $a = \ln(SUM2) - (b)(COC2)$

   where

   COC1 = low concentration of concern
   COC2 = high concentration of concern
   SUM1 = sum of probabilities of COC1
   SUM2 = sum of probabilities of COC2

2. The following equation is used to estimate the missing sum of probabilities:

   $$SUM = \exp[a + b\ (COC)]$$

   where

   COC is the ratio of concentration of concern to loading entered by the user that is not represented in the matrix file.

The piece-wise exponential interpolation method was tested by comparing some interpolated values of the sum of probabilities with their actual values. Using the mainframe computer program, some actual values of the sum of probabilities that are not represented in the matrix file by SIC code group 2865 were obtained. Table 6-2 shows the actual and interpolated values at 6 concern level/loading ratios. Also presented in Table 6-2 are the relative percentages of discrepancies between the actual and the interpolated values.

6.4.3 PC Program

The PC program to produce probability results is written in Turbo Pascal®. The four major parts of the program are as follows:

1. Prompts to the user for SIC group, number of days of release, number of sites (plants), loading rate, and concentration of concern levels;
2. Acquisition of appropriate SIC probability matrix file;
3. Interpolation of the probability matrix file; and
4. Calculation of the final results.

A copy of the program is found in the programmer's documentation. The fourth part is discussed in the next section.

Each of the phases was checked by running the program for a given concern level and loading rate and comparing the results to hand calculated results to assure correct programming.

6.5 Option 2 Results

The final results of Option 2 for each concentration of concern level entered are:

1. Times per year the concentration of concern level is exceeded and
2. Percent of year the concern level is exceeded.

Table 6-2. Test Reslts of Interpolation Method

Concern level/ loading ratio	Actual sum of probabilities	Interpolated sum of probabilities	Relative percentages of discrepancy
0.00023	265.11377	265.11625	0.0009%
0.01030	223.64680	223.65036	0.0015%
0.30300	143.25670	143.25750	0.0005%
1.03000	100.02038	100.02663	0.0062%
30.30000	9.70139	9.70166	0.0027%
1030.00000	0.06047	0.60900	0.04961%

The first result is calculated as follows:

$$\text{Times/Yr Conc. Exceeded} = (ND)\,\frac{SUM}{NFS}$$

where

ND = number of release days for the facility
SUM = sum of the 10th percentile probability of exceedance
NFS = number of facilities in SIC group used to calculate the sum of probabilities (Table 6-1).

The second result is simply the percentage of the first and is calculated as follows:

$$\%\text{ of Year Conc. Exceeded} = \frac{(\text{Times/yr. conc. exceeded})(100)}{365}$$

7. REFERENCES

Abramowitz M, Stegun IA. 1954. Handbook of Mathematical Functions. National Bureau of Standards Applied Mathematics Series 55. Washington DC: Superintendant of Documents. U.S. Government Printing Office.

DiToro DM. 1984. Probability model of stream quality due to runoff. ASCE. Journal of Environmental Engineering. Vol 110(3):607-628.

USEPA. 1984. U.S. Environmental Protection Agency. Technical Guidance Manual for Performing Waste Load Allocations. Book VII. Permit Averaging Permits. Washington, DC: Office of Water Regulations and Standards.

Versar. 1984. A regression analysis of stream flow data stratified by USGS major basin. Washington, DC: Office of Toxic Substances, U.S. Environmental Protection Agency. Contract No. 68-02-3968.